Journal of Machine Learning Research 17 (2016) 1-37 Submitted 3/15; Revised 1/16; Published 11/16

Machine Learning in an Auction Environment

Patrick Hummel PHUMMEL@GOOGLE.COM
1600 Amphitheatre Parkway
Google Inc.
Mountain View, CA 94043, USA

R. Preston McAfee PRESTON@MCAFEE.CC
One Microsoft Way
Microsoft Corp.
Redmond, WA 98052, USA

Editor: Shie Mannor

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 in an auction environment is exceedingly small.

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 way to determine 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 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. 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 different auctions. In these settings, there will always be a certain amount of free exploration that takes place due to the fact that there will be some auctions in which there 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 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,[1] the value of learning in the problem that we consider is dramatically different from the value of learning in standard multi-armed bandit problems. In standard multi-armed bandit problems (Auer et al., 2002) where there is no discounting of payoffs and no random variation in the competition that an arm faces, 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 would be zero.

Our baseline model considers a simple situation in which there is 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. We show that our conclusions about the optimal bidding strategies extend to a variety of more complicated models including models in which there are multiple advertisers with unknown eCPMs 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. We also illustrate an asymptotic equivalence between

1. In addition, Iyer et al. (2014) illustrate that bidders may have an incentive to make a bid equal to their expected value plus a term proportional to their value of learning about their value if bidders have uncertainty about their own value. This paper differs from ours in that it considers an environment in which bidders are attempting to learn their own values rather than an auctioneer attempting to learn the eCPMs.

the theoretically optimal strategies and the strategies that would be selected by a simple one-step look ahead policy often referred to as "knowledge gradients" (Frazier et al., 2009; Ryzhov et al., 2010, 2012).

A consequence of these small incremental changes in the probability that an ad is shown is that the total value from adding active exploration 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 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, the maximum total efficiency gain that could ever be achieved by adding active learning 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 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 a paper by Li et al. (2010). This paper is the only other paper we are aware of that considers questions related to the value of learning about the eCPMs 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. Li et al. (2010) demonstrate 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, Li et al. (2010) 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 (Audibert and Bubeck, 2010; Auer et al., 2002, 2003; Gittins, 1979; Hazan and Kale, 2011; Lai and Robbins, 1985; Mannor and Tsitsiklis, 2004; May et al., 2012; Slivkins, 2014) as well as some papers that focus specifically on the auction context (Agarwal et al., 2009; Babaioff et al., 2009; Devanur and Kakade, 2009; Wortman et al., 2007). However, none of these papers considers appropriate methods for exploring ads in a context where there is random variation in the quality of the competition that an ad faces in an auction. The optimal methods for exploring ads in such a scenario turn out to be completely different from 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 in 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 (Bergemann and Välimäki, 1996, 1997, 2000), firms and sellers trying to learn about demand (Aghion et al., 1993; Fishman and Rob, 1998; Ghate, 2015; Keller and Rady, 1999; Mirman et al., 1993; Rusitchini and Wolinsky, 1995), learning to play repeated games (Anthonisen, 2002; Gale and Rosenthal, 1999), learning about untried policies in political economy (Callander, 2011; Callander and Hummel, 2014; Strulovici, 2010), learning from the actions of others (Banerjee and Fudenberg, 2004; Gale, 1996; Vives, 1997), as well as general results on experimentation

(Aghion et al., 1991; Banks and Sundaram, 1992; Bergemann and Välimäki, 2001; Bolton and Harris, 1999; Brezzi and Lai, 2002; Keller and Rady, 2010; Keller et al., 2005; Moscarini and Smith, 2001; Rothschild, 1974; Schlag, 1998; Weitzman, 1979). However, the economics literature has not considered strategic experimentation in auctions, as we do in the present paper.

2. The Model

There is a new ad with an uncertain eCPM that will bid into a second-price auction for a single advertising opportunity with competing advertisers.[2] Throughout we let x denote the actual unknown but fixed value (or eCPM) for showing the new ad, z denote the eCPM bid the auctioneer places on behalf of this advertiser,[3] and let k denote the number of impressions the ad has received so far. We also suppose that the highest eCPM bid that this advertiser competes against may vary from auction to auction, and that in each auction, this highest competing eCPM bid is a random draw from some cumulative distribution function $F(\cdot)$ with corresponding continuous and twice differentiable density $f(\cdot)$.

At any given point in time, the auctioneer does not necessarily know the exact value of x. Instead the auctioneer only knows that x is drawn from some distribution. We let $\tilde{x}$ denote a generic distribution corresponding to the auctioneer's estimate of the distribution of possible values of x. 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.

Throughout we also let $\overline{x}$ denote an unbiased estimate of the true value of x given the auctioneer's estimate of the distribution of possible values of x. 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 $\frac{s^2(\overline{x})}{k}$ for some constant $s^2(\overline{x})$ that depends only on $\overline{x}$, and we assume that $\sigma_k^2 = \frac{s^2(\overline{x})}{k} + \frac{h(\overline{x})}{k^2} + o(\frac{1}{k^2})$ for some continuously differentiable functions $s^2(\overline{x})$ and $h(\overline{x})$.[4]

In addition, we let $\delta \in (0,1)$ denote the per-period discount rate so the auctioneer 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. Throughout we assume that the auctioneer wishes to maximize total efficiency; that is, if v_t denotes the total value of the ad displayed in period t (the true eCPM of this ad), then the auctioneer's payoff is $\sum_{t=0}^{\infty} \delta^t v_t$. Since online ad auctions are typically designed to select the efficiency-maximizing allocation, this is a logical objective to optimize.

2. These second-price auctions for a single advertising slot are ubiquitous throughout the display advertising industry. In such auctions, advertisers have an incentive to make a bid equal to their true value for a click.
3. Typically advertisers bid by indicating how much they are willing to pay per click, and the auctioneer then uses this cost-per-click bid as well as an estimate of the probability the ad will be clicked to calculate an eCPM bid for the advertiser that the auctioneer then places on behalf of the advertiser in the auction.
4. This assumption will hold for most common priors about the distribution from which the uncertain eCPM of the ad is drawn, such as a beta prior. It is also worth noting that the weaker assumption that $\sigma_k^2 = \frac{s^2(\overline{x})}{k} + O(\frac{1}{k^2})$ for some constant $s^2(\overline{x})$ is sufficient to prove our main result about the value of learning being $O(\frac{1}{k^2})$. The additional assumption that $\sigma_k^2 = \frac{s^2(\overline{x})}{k} + \frac{h(\overline{x})}{k^2} + o(\frac{1}{k^2})$ is only used to further prove that the value of learning is of the form $\frac{v(\overline{x})}{k(k+1)} + o(\frac{1}{k^2})$ for some function $v(\overline{x})$.

3. Preliminaries

Before proceeding to analyze the precise model given above, we first address a closely related question about the extent to which a particular advertising opportunity increases total welfare if the eCPM of this advertising opportunity is known. In particular, we consider a concept that we refer to as the *long-term value* of a particular advertisement. The long-term value of a particular advertisement gives the total increase in the auctioneer's payoff that arises as a result of this ad being in the system from the various auctions that take place over time. Understanding the long-term value of a particular advertisement when the eCPM of that ad is known will serve as a useful benchmark for understanding how one should behave when there is uncertainty about the eCPM of the ad.

Theorem 1 *If the eCPM of an ad is known, then the total long-term value of this ad is a convex function of the eCPM of the ad and a strictly convex function for regions where the eCPM of the ad is within the support of the distribution of the highest competing eCPM.*

All proofs are in the appendix. The fact that the value for any particular advertisement is a convex function of the eCPM of the ad if the eCPM of the ad is known indicates that if there is uncertainty about the eCPM of the ad, then the expected long-term value of this ad will be greater than the long-term value of the expected eCPM of the ad. From this it follows that if there is uncertainty about the eCPM of the ad, then it will be optimal to behave as if this particular ad had a known eCPM that is greater than the expected eCPM of the ad. The precise additional amount that this advertiser's bid should be increased will be pinned down by the solution to the dynamic programming problem governed by the game described in the model.

4. 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. First we derive the auctioneer's payoff that arises in a particular period when the auctioneer makes a particular bid on behalf of the advertiser with uncertain eCPM.

Note that if the auctioneer places a bid of z on behalf of the advertiser with uncertain eCPM in the auction and the actual value of showing this particular ad is x, then the auctioneer's payoff from running the auction once is

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

In general placing a bid of z rather than x in a one-shot auction will result in some inefficiencies in the one-shot auction since it would be optimal for efficiency to place a bid exactly equal to x on behalf of this advertiser in a one-shot auction. The payoff loss that arises in a one-shot auction as a result of placing a bid of z instead of x is

$$L = u(x,x) - u(z,x)$$

$$= x(1-F(x)) + \int_x^\infty (1-F(y))\,dy + xF(x) - z(1-F(z)) - \int_z^\infty (1-F(y))\,dy - xF(z)$$
$$= (x-z)(1-F(z)) + \int_x^z (1-F(y))\,dy = \int_x^z F(z) - F(y)\,dy.$$

If we define the per-period reward to be the negative of this per-period loss, then the auctioneer seeks to maximize the discounted sum of these per period rewards. Let $V_k(\overline{x})$ denote the value of this discounted sum when the auctioneer follows the optimal bidding strategy. Also note that, from the perspective of the auctioneer, x is a random variable that can be expressed as $x = \overline{x} + \sigma_k \epsilon$, where σ_k denotes the standard deviation in our estimate of the ad with uncertain eCPM when the ad has been shown k times, and ϵ is a random variable with mean zero and variance one. We use this notation to prove the following:

Lemma 2 *$V_k(\overline{x})$ can be expressed as the value of a dynamic programming problem by*

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

where $\overline{x}'$ denotes the uncertain realization of $\overline{x}$ after an ad receives an additional impression.

By using the expression for the value of the dynamic programming problem in the previous lemma, we can derive the bid that the auctioneer should place on behalf of the advertiser to maximize the auctioneer's payoff. This is done in the theorem below:

Theorem 3 *The optimal bidding strategy in the dynamic programming problem when an ad has been shown k times entails setting $z = \overline{x} + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$.*

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

5. Value of Dynamic Program for Large Numbers of Impressions

In the previous section, we have given exact expressions for the value of the dynamic program 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 program in the limit when an ad has already been shown a large number of times.

The main purpose of this section is to illustrate that the value of learning term given in the previous section will vary with $\frac{1}{k^2}$ for large k. We prove this by first showing that the expected efficiency loss arising due to the uncertainty in the eCPM of the ad varies with $\frac{1}{k}$ for large k, and then use this to show that the value of learning term varies with $\frac{1}{k} - \frac{1}{k+1}$, which varies with $\frac{1}{k^2}$ for large k.

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:

Lemma 4 *$E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\, dy] = \int_{\overline{x}}^{z} F(z) - F(y)\, dy + \frac{1}{2}\sigma_k^2 f(\overline{x}) + a(\overline{x})\sigma_k^4 + o(\sigma_k^4)$ for some constant $a(\overline{x})$ for large k.*

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

Theorem 5 *$V_k(\overline{x}) = \Theta(\frac{1}{k})$ 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 5.

Theorem 5 suggests that we may be able to write $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + o(\frac{1}{k})$ for large k, where v is a function that depends only on $\overline{x}$. To prove that $V_k(\overline{x})$ can be expressed this way, it is necessary to show that $kV_k(\overline{x})$ indeed converges to a function of $\overline{x}$ in the limit as $k \to \infty$. This is done in the following theorem:

Theorem 6 *$kV_k(\overline{x})$ converges to a function of $\overline{x}$ in the limit as $k \to \infty$. Furthermore, it must be the case that $kV_k(\overline{x}) = -\frac{1}{2(1-\delta)} s^2(\overline{x}) f(\overline{x}) + O(\frac{1}{k})$ for large k.*

From Theorem 6, it follows that we can express $V_k(\overline{x})$ by $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + O(\frac{1}{k^2})$ for large k, where v is a function that satisfies $v(\overline{x}) = \frac{1}{2(1-\delta)} s^2(\overline{x}) f(\overline{x})$. In order to complete our approximation of the solution the dynamic programming problem for large k, it is also necessary to bound the expression $E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})$ that appears in the dynamic programming problem. This is done in the following theorem:

Theorem 7 *$E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k.*

The intuition behind this result is that since the efficiency loss that the auctioneer incurs due to 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 $E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})$ 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 implies that the incremental increase in an advertiser's probability of winning the auction will also vary with $\frac{1}{k^2}$ for large k.

The result in Theorem 7 suggests that the optimal method for adding active exploration 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 arm faces

in a given period (Auer et al., 2002). 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.

Ultimately we seek to use these insights to derive results about the change in payoff that would result from incorporating active learning in this setting. Before doing this, we first illustrate how the conclusions of this section about the value of the dynamic programming problem and the optimal bidding strategy extend to a variety of more complicated scenarios including settings where there are multiple different ads with uncertain eCPMs whose true eCPMs may be correlated and we also illustrate a natural correspondence between the optimal solution to the full dynamic programming problem and a simple one-step look-ahead strategy. First we tackle the problem of computing the value of the dynamic program when an ad with an uncertain eCPM has only received a small number of impressions.

6. Value of Dynamic Program for Small Numbers of Impressions

To calculate the value of $V_k(\overline{x})$ for small values of k, we apply backwards induction. At some large value of k, it will necessarily be the case that the incremental value of additional exploration is so small that the advertiser simply bids $z = \overline{x}$ because the smallest possible increment the advertiser would be allowed to adjust its bid exceeds the tiny incremental value of additional exploration. Thus if K denotes the earliest stage at which an advertiser always sets $z = \overline{x}$, then for all $k \geq K$, it is necessarily the case that the value of learning is zero, and $V_k(\overline{x}) = \frac{1}{1-\delta}\left(E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{\overline{x}} F(\overline{x}) - F(y)\ dy\right]\right) \approx 0$.

For values of $k < K$, we have

$$(1-\delta)V_k(\overline{x}) = E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z_k} F(z_k) - F(y)\ dy\right] + \delta F(z_k)(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$$

or

$$V_k(\overline{x}) = \frac{E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z_k} F(z_k) - F(y)\ dy\right] + \delta F(z_k) E_{\overline{x}'}[V_{k+1}(\overline{x}')]}{1-\delta+\delta F(z_k)}.$$

Thus by empirically measuring the values of σ_k and $F(\cdot)$, we can apply backward induction to approximate $V_k(\overline{x})$ for small values of k. We now address the question of what these values of $V_k(\overline{x})$ will be approximately equal to for an important class of advertisers.

Many ads that have only received a small number of impressions are ads that typically fail to win auctions because the machine learning system is pessimistic about the ad's true eCPM. The estimated eCPMs for these ads may be several orders of magnitude smaller than the typical eCPMs of the ads that have been shown many times. In these cases, even if the percentage uncertainty in the eCPMs of these ads is quite high, the absolute amount of uncertainty in the eCPMs of these ads will be small compared to the typical eCPMs of the ads that have been shown many times. Thus in these cases, $\overline{x}$ will be close to zero, and $F(\overline{x})$ and σ_k^2 will be close to zero as well. Under these circumstances, we have the following result:

Theorem 8 *If $\overline{x}$ (and σ_k^2) are close to zero for small values of k, then $V_k(\overline{x}) = -\frac{1}{2(1-\delta)} f(\overline{x})\sigma_k^2 + o(f(\overline{x})\sigma_k^2)$ for small values of k.*

Theorem 8 indicates that even in small sample environments, it is still frequently reasonable to approximate $V_k(\overline{x})$ by writing $V_k(\overline{x}) \approx -\frac{1}{2(1-\delta)} f(\overline{x})\sigma_k^2$, where σ_k^2 denotes the variance in our estimate of the ad's eCPM for a particular value of k. This theorem in turn implies that if a machine learning system is quite pessimistic about the true eCPM of a new ad, then there will be little value to actively exploring the ad because the value of learning term, $E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})$, will be quite small.

7. Ads with Correlated Values

So far we have restricted attention settings in which we only seek to learn the eCPM of one advertiser's ad. However, in many situations we may seek to learn the eCPMs of multiple advertisers' ads and the eCPMs of the various advertisers may be correlated. In these situations, information about one ad's eCPM may help one learn about the eCPMs of other related advertisers. On top of this, even if there is only ad for which we are uncertain about the advertiser's eCPM, this ad may bid in several different contexts where the ad has substantially different eCPMs and the ad faces substantially different competing landscapes of bids.[5] In these cases, information about an ad's eCPM in one context may also help one learn about the ad's eCPM in other contexts.

To address how this affects the results, we extend the model to allow for the possibility that there are multiple different ads that bid in multiple different contexts where we seek to learn the eCPMs of the ads and these eCPMs may be correlated. In particular, we suppose that there are m different ad-context pairs where we seek to learn the eCPM of the ad in that particular context. For the ad-context pair a, we let x_a denote the actual, unknown value of the eCPM of that ad in that context, and we let $x = (x_1, \ldots, x_m)$ denote the actual unknown eCPMs of the ads in all m contexts. We also let k denote the total number of impressions that these advertisers have received in the various contexts and let β_a denote the fraction of these impressions that were received in context a. Thus we have $\sum_{a=1}^{m} \beta_a = 1$.

We again assume the auctioneer does not know the exact value of x, and instead the auctioneer only knows that x is drawn from some distribution. We again let $\tilde{x}$ denote a generic distribution corresponding to the auctioneer's estimate of the distribution of possible values of x. This distribution allows for the possibility that the auctioneer may believe there is correlation in the unknown eCPMs of the advertisers in the different contexts, and the distribution will again evolve over time as an ad has received more impressions and we have a better sense of the underlying eCPM of the ad.

Throughout we also let $\overline{x}$ denote an unbiased estimate of the true value of x given the auctioneer's estimate of the distribution of possible values of x and we let $\overline{x}_a$ denote an unbiased estimate of the true value of x_a given this distribution. We also let σ^2_{a,k_a} denote the variance in our estimate of the eCPM of the ad-context pair a when there have been a total of k_a impressions in this ad-context pair. In the limit when k_a is large, σ^2_{a,k_a} will be well approximated by $\frac{s_a^2(\overline{x}_a)}{k_a} = \frac{s_a^2(\overline{x}_a)}{\beta_a k}$ for some constant $s_a^2(\overline{x}_a)$ that depends only on $\overline{x}_a$, and

5. Contextual bandit problems in which an arm's payoff may vary from context to context have appeared in the literature before in different settings. See, for example, work by May et al. (2012) and Slivkins (2014).

we again let $\sigma^2_{a,k_a} = \frac{s^2_a(\overline{x}_a)}{\beta_a k} + \frac{h_a(\overline{x}_a)}{\beta_a^2 k^2} + o(\frac{1}{k^2})$ for some continuously differentiable functions $s^2_a(\overline{x}_a)$ and $h_a(\overline{x}_a)$. 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 each period t, there is an auction for a single advertising opportunity. The auction can involve any one of the m possible ad-context pairs for which we do not know the eCPM of the ad in that context. We let π_a denote the probability that there will be an auction involving ad-context pair a in any given time period. Thus we have $\sum_{a=1}^m \pi_a = 1$. We further suppose that if there is an auction involving ad-context pair a, then 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 $F_a(\cdot)$ with corresponding continuous and twice differentiable density $f_a(\cdot)$.

In this setting, the total long-term value of a particular ad-context pair is again a convex function of the eCPM of the ad for the same reasons as in Theorem 1 and we can again formulate this problem as a dynamic program. To do this, let $\vec{k} \equiv (k_1, \ldots, k_m)$ denote a vector that gives the number of impressions that have been received by the various ad-context pairs $1, \ldots, m$. Also let $V_{a,\vec{k}}(\overline{x})$ denote the value of the dynamic program when the next auction involves the advertiser-context pair a, the eCPMs of the ads are $\overline{x}$, and there have been $\vec{k}$ impressions in each of the various ad-context pairs, and let $V_{\vec{k}}(\overline{x}) \equiv E_a[V_{a,\vec{k}}(\overline{x})]$ denote the value of the same dynamic program unconditional on which ad-context pair is involved in the next auction. By using similar reasoning to that in Lemma 2, we know that $V_{\vec{k}}(\overline{x})$ equals

$$\frac{1}{1-\delta} E_a \left[\max_{z_a} E_\epsilon \left[-\int_{\overline{x}_a + \sigma_{a,k_a}\epsilon}^{z_a} F_a(z_a) - F_a(y)\, dy + \delta F_a(z_a)(E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x})) \right] \right],$$

where $\vec{k}'(a,\vec{k}) \equiv (k'_1, \ldots, k'_m)$ is a vector that satisfies $k'_b = k_b$ for all $b \neq a$ and $k'_a = k_a + 1$, and $\overline{x}'(a)$ denotes the uncertain realization of $\overline{x}$ if the advertiser-context pair a receives an additional impression. Furthermore, the optimal bid z_a if there is an auction involving the advertiser-context pair a satisfies $z_a = \overline{x}_a + \delta F_a(z_a)(E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x}))$ by similar logic to that given in Theorem 3, and the result in Lemma 4 is just a general mathematical result that holds regardless of the model we are considering. Thus natural analogs of Theorems 1 and 3 and Lemmas 2 and 4 continue to hold in this revised model.

By using these insights, one can further show that $V_{\vec{k}}(\overline{x})$ must be on the order of $\frac{1}{k}$ for large k. This is done in the following theorem:

Theorem 9 *When there are multiple ads with correlated values, $V_{\vec{k}}(\overline{x}) = \Theta(\frac{1}{k})$.*

While this result indicates that $V_{\vec{k}}(\overline{x})$ varies with $\frac{1}{k}$ for large values of k, this alone does not guarantee the convergence of this function for large values of k. We verify that this function does indeed converge for large values of k in the following theorem:

Theorem 10 *When there are multiple ads with correlated values, $kV_{\vec{k}}(\overline{x})$ converges to a function of $\overline{x}$ in the limit as $k \to \infty$. Furthermore, it must be the case that $kV_{\vec{k}}(\overline{x}) = -\frac{1}{2(1-\delta)} \sum_{a=1}^m \pi_a \frac{1}{\beta_a} s^2_a(\overline{x}_a) f_a(\overline{x}_a) + O(\frac{1}{k})$ for large k.*

Theorem 10 indicates that the results about the limiting value of $kV_{\vec{k}}(\overline{x})$ derived in Theorem 6 extend naturally to the case where there are multiple ads with possibly correlated values. When there are multiple ad-context pairs that we must learn about, the value function corresponding to that in Theorem 6 differs only in that we take a weighted sum over the various possible advertiser-context pairs, where the weights are a function of the relative probabilities with which each advertiser-context pair arises. Thus there is a clear analog between the limiting properties of the value function when there are multiple advertiser-context pairs and the value function in the main model.

By using the results in the previous theorem, one can further derive properties of the limiting value of $E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x})$ that is proportional to the additional amount that one should bid in the auction beyond the expected value that one has for the advertising opportunity. This is stated below in the following theorem:

Theorem 11 *When there are multiple ads with correlated values,* $E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x}) = -\frac{\pi_a}{2(1-\delta)\beta_a^2 k^2} s_a^2(\overline{x}_a) f_a(\overline{x}_a) + o(\frac{1}{k^2})$ *for large* k.

The proof of this result is substantively identical to the proof of Theorem 7 and is thus omitted. Theorem 11 illustrates that the substantive conclusions of Theorem 7 extend to this alternative environment in which there are multiple ads with possibly correlated values. When there are multiple ads, it remains optimal to increase one's bid by an amount proportional to the variances in our estimates of the eCPMs of the ads or $\frac{1}{k^2}$ for large k.

8. Knowledge Gradients

Throughout the paper so far, we have considered a standard dynamic programming approach in which the optimal decision at any given point in time is affected in part by how this decision will affect future decisions when looking at the infinite horizon ahead. While this is a standard approach to take in these types of problems, recently there has been work considering an alternative approach often referred to as "knowledge gradients" in which the decision one takes in a given period is the decision that one would take if one faced an infinite-horizon game but this period was the last period in which the information one learned could be used to inform future actions.

The main advantage of these knowledge gradients over the standard dynamic programming approach is that they have the virtue of being much easier to calculate than the optimal bidding strategy under the standard dynamic programming problem. This simplicity does potentially come at a performance cost. However, various papers have illustrated that using this simple one-step look-ahead approach can nonetheless achieve a performance that is competitive with that of other standard methods in contexts unrelated to advertising (Frazier et al., 2009; Ryzhov et al., 2010, 2012). In this section, we investigate whether this alternative knowledge gradient approach can indeed achieve a performance comparable to that of the theoretically optimal dynamic programming approach.

To address this question, for simplicity we consider the baseline model in which there is one advertisement for which we are seeking to learn the eCPM of the ad, though similar results can easily be derived under the more general model we have considered with multiple ads and correlated values. Let $U_k(\overline{x})$ denote the value that one would obtain for the rest of

the game when an ad has received k impressions so far, one's estimate of the eCPM of the ad is $\overline{x}$, and one will not be able to use information that one learns in the future to inform future bidding decisions. Note that in this case, the optimal bidding strategy will be to submit a bid of $z = \overline{x}$ in every remaining period, and the auctioneer's expected per-period payoff will be $E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy\right]$ in every future period, where ϵ denotes some random variable with mean zero and variance one, and $z \equiv \overline{x}$. The total value the auctioneer will obtain for the rest of the game is then $U_k(\overline{x}) = \frac{1}{1-\delta}E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{\overline{x}} F(\overline{x}) - F(y)\ dy\right]$.

Also let $U_{k+1}(\overline{x})$ denote the value that one would obtain for the rest of the game when an ad has received $k+1$ impressions so far, one's estimate of the eCPM of the ad is $\overline{x}$, and one will not be able to use information that one learns in the future to inform future bidding decisions. The total value the auctioneer will obtain for the rest of the game is then $U_{k+1}(\overline{x}) = \frac{1}{1-\delta}E_\epsilon\left[-\int_{\overline{x}+\sigma_{k+1}\epsilon}^{\overline{x}} F(\overline{x}) - F(y)\ dy\right]$.

Now consider the bidding strategy that one would employ if one faced an infinite-horizon game but this period was the last period in which the information one learned could be used to inform future actions. The auctioneer's payoff from bidding z that arises in the current period equals $-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy$. And the expected value that the auctioneer obtains from future periods by bidding z in the current period is $F(z)E_{\overline{x}'}[U_{k+1}(\overline{x}')] + (1-F(z))U_k(\overline{x})$ by the same reasoning used in the proof of Lemma 2. From this it follows that the expected payoff from placing a bid of z in a given period is

$$E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy + \delta(F(z)E_{\overline{x}'}[U_{k+1}(\overline{x}')] + (1-F(z))U_k(\overline{x}))\right].$$

There is a clear similarity between this expression and the expression for the expected payoff from placing a bid of z in the standard dynamic programming approach. The main difference is that the terms $U_{k+1}(\overline{x}')$ and $U_k(\overline{x})$ have replaced the terms $V_{k+1}(\overline{x}')$ and $V_k(\overline{x})$ in the standard dynamic programming approach.

It is worth noting, however, that the payoffs that result from the one-step look ahead strategies in the model in this paper take a different form than those given in other knowledge gradient papers (Frazier et al., 2009; Ryzhov et al., 2010, 2012). The reason for this difference is that in the model in our paper, there is a competing ad whose eCPM is known in each period but is also a random draw from some distribution in each period. No such random changes in the values of the arms from period to period are present in existing knowledge gradient papers, so the payoffs and strategies in our paper are formulated differently than those given in existing knowledge gradient papers.

From the equation we've derived for the auctioneer's payoff from bidding z, we can calculate the optimal bidding strategy under the knowledge gradient formulation. This bidding strategy is given in the following theorem:

Theorem 12 *The optimal bidding strategy in the knowledge gradient framework when an ad has been shown k times entails setting $z = \overline{x} + \delta(E_{\overline{x}'}[U_{k+1}(\overline{x}')] - U_k(\overline{x}))$.*

The proof of this result is substantively identical to that in Theorem 3 and is thus omitted. This result indicates that in the knowledge gradient framework, the incremental amount that one increases one's bid beyond the immediate expected reward is again of the

form $\delta(E_{\overline{x}'}[U_{k+1}(\overline{x}')] - U_k(\overline{x}))$, the only difference being that $U_k(\overline{x})$ corresponds to the value of the dynamic program under the knowledge gradient framework.

To better understand the incremental amount that one would increase one's bid, we present two results that illustrate how the incremental amount that one would increase one's bid under the knowledge gradient framework compares to the incremental amount that one would increase one's bid under the full dynamic programming problem. First we present a finite sample result about how these incremental bid increases compare in the two frameworks.

In our first result, we consider what we refer to as the expected value of all future learning. To reflect the fact that $V_k(\overline{x})$ gives the auctioneer's payoff from the full dynamic programming problem when the auctioneer is able to make use of additional information in future periods, whereas $U_k(\overline{x})$ gives the auctioneer's payoff from the corresponding game in which the auctioneer is not able to make use of information that he learns, we define this expected value of all future learning term to be the difference between $V_k(\overline{x})$ and $U_k(\overline{x})$. With this definition in mind, we obtain the following result:

Theorem 13 *Suppose the expected value of all future learning is lower after the ad has been shown $k+1$ times than it is after the ad has been shown k times. Then the incremental amount by which one would increase one's bid under the knowledge gradient framework is greater than it is under the full dynamic programming problem.*

Theorem 13 indicates that the solution to the one-step look ahead problem will generally involve increasing one's bid beyond the immediate expected value of the advertising opportunity by a greater amount than one would do so under the full dynamic programming problem. This makes sense intuitively. If the current period were the last period in which one could ever use information that one learns to inform future actions, then one would place quite a high premium on being able to learn this information while one still can. By contrast, in the full dynamic programming problem, there will always be plenty of opportunities to learn this information later, so there is relatively less incentive to substantially increase one's bid beyond the immediate expected reward. This explains the result in Theorem 13.

Theorem 13 requires a technical condition that the expected value of all future learning is lower if an ad has been shown $k+1$ times than if the ad has been shown k times, but this is just a mild technical constraint that we would expect to hold in virtually any situation. When an ad has been shown $k+1$ times, one has more precise information about the true eCPM of the ad than when the ad has only been shown k times, so there is less value to learning more about the true eCPM of the ad.

While Theorem 13 suggests that one might increase one's bid by too much under the knowledge gradient framework compared to the strategy that one should follow under the full dynamic programming problem, these differences in bidding strategies turn out to be relatively small. We illustrate this by characterizing the value of $E_{\overline{x}'}[U_{k+1}(\overline{x}')] - U_k(\overline{x})$:

Theorem 14 *In the knowledge gradient framework, $E_{\overline{x}'}[U_{k+1}(\overline{x}')] - U_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k, where $v(\overline{x}) \equiv \frac{1}{2(1-\delta)} s^2(\overline{x}) f(\overline{x})$.*

This result also immediately implies the following corollary:

Corollary 15 *The ratio of the incremental amount by which one wants to increase one's bid in the knowledge gradient framework and the incremental amount by which one wants to increase one's bid in the standard dynamic programming approach becomes arbitrarily close to* 1 *in the limit as the amount of uncertainty in an ad's eCPM becomes arbitrarily small.*

Thus using the knowledge gradient to formulate one's bidding strategy will result in payoffs that are asymptotically equivalent to those that would result from using the theoretically optimal bidding strategy. These results suggest that the knowledge gradient framework is indeed an appealing framework for computing bidding strategies in an environment where one wishes to learn about the unknown eCPMs of advertisers, as using this approach will result in little loss from the theoretically optimal approach.

9. Learning About Multiple Advertisers in the Same Auction

In the analysis so far, we have assumed that in any given auction, there is only one advertiser whose eCPM is unknown. But in many real-life auctions there may be multiple advertisers with unknown eCPMs. In these cases, an auctioneer must decide both which advertiser with unknown eCPM will have the highest bid as well as what bid to submit for this advertiser.

In this setting, it is not clear whether the decision maker's optimal strategy can simply be represented by submitting a bid for each advertiser that is equal to the sum of the best estimate of the advertiser's eCPM as well as a value of learning term. It may be the case that the optimal bid for advertiser i if advertiser i submits the highest bid of the advertisers with unknown eCPMs is higher than the optimal bid for some other advertiser j if advertiser j submits the highest bid of the advertisers with unknown eCPMs, even though the decision maker would prefer to submit a higher bid for advertiser j than for advertiser i. We address whether this possibility can arise in this section.

To address this question, suppose that in each auction, there are n ads with unknown eCPMs. The actual eCPMs of these ads are $x_1, \ldots, x_n$, and we let $z_1, \ldots, z_n$ denote the bids placed by these advertisers in the auction. Also let i denote the advertiser who submits the highest eCPM bid amongst these n bidders and let j denote the advertiser who actually has the highest eCPM amongst these advertisers. In each auction, these advertisers with unknown eCPMs compete against other advertisers and the highest such competing eCPM is drawn from a cumulative distribution function $F(\cdot)$ with corresponding density $f(\cdot)$.

Note that in this case, the utility that the decision maker obtains in a given period from having advertiser i submit a bid of z_i that is the highest eCPM bid amongst these n bidders is $u = z_i(1 - F(z_i)) + \int_{z_i}^{\infty}(1 - F(y))\, dy + x_i F(z_i)$. At the same time, this decision maker would obtain a utility of $u = x_j(1 - F(x_j)) + \int_{x_j}^{\infty}(1 - F(y))\, dy + x_j F(x_j)$ in a given period from making the optimal decision in a given period. Thus the loss that this decision maker obtains in a given period as a result of having advertiser i submit a bid of z_i that is the highest eCPM bid amongst these n bidders is the difference between these two utilities or $L = x_j - z_i + (z_i - x_i)F(z_i) + \int_{x_j}^{z_i}(1 - F(y))\, dy = \int_{x_i}^{z_i} F(z_i)\, dy - \int_{x_j}^{z_i} F(y)\, dy$.

Now let k_i denote the number of impressions that advertiser i has received so far, let $\vec{k} \equiv (k_1, \ldots, k_n)$ denote a vector that gives the number of times each of these ads has been shown, let $\overline{x}_i$ denote our best estimate of the expected value of the eCPM of advertiser i,

and let $\overline{x} \equiv (\overline{x}_1, \ldots, \overline{x}_n)$ denote a vector of these best estimates. Also let $V_{\vec{k}}(\overline{x})$ denote the value of the dynamic program as a function of these quantities.

By similar reasoning to that in the proof of Lemma 2, it follows that if i denotes the advertiser who submits the highest eCPM bid amongst the n bidders with unknown eCPMs, $\vec{k}'(i) \equiv (k'_1, \ldots, k'_n)$ is the vector where $k'_j = k_j$ for all $j \neq i$ and $k'_i = k_i + 1$, and $\overline{x}'(i)$ denotes the uncertain realization of $\overline{x}$ if advertiser i receives an additional impression, then $V_{\vec{k}}(\overline{x})$ equals

$$\frac{1}{1-\delta}\left(\max_{z_i} E_{x_i,x_j}\left[\int_{x_j}^{z_i} F(y)\, dy - \int_{x_i}^{z_i} F(z_i)\, dy + \delta F(z_i)(E_{\overline{x}'(i)}[V_{\vec{k}'(i)}(\overline{x}'(i))] - V_{\vec{k}}(\overline{x}))\right]\right),$$

where the difference in the values of the dynamic programs is due to the fact that the loss in a given period is now $\int_{x_i}^{z_i} F(z_i)\, dy - \int_{x_j}^{z_i} F(y)\, dy$. Similarly, the optimal bid for advertiser i if advertiser i submits the highest eCPM bid amongst the n bidders with unknown eCPMs still satisfies $z_i = \overline{x}_i + \delta(E_{\overline{x}'(i)}[V_{\vec{k}'(i)}(\overline{x}'(i))] - V_{\vec{k}}(\overline{x}))$. We use these insights to prove the following:

Theorem 16 *Suppose the optimal bid for advertiser i if advertiser i submits the highest bid of the advertisers with unknown eCPMs is higher than the optimal bid for all other advertisers with unknown eCPMs if one of these other advertisers submits the highest bid of the advertisers with unknown eCPMs. Then it is also optimal for advertiser i to have the highest bid of all the advertisers with unknown eCPMs.*

Theorem 16 guarantees that if there are multiple ads with unknown eCPMs, then one can simply compute the optimal bids for each of these ads in the case where the ad in question was guaranteed to have a higher bid than the other ads with unknown eCPMs. The ad that has the highest such optimal bid will then be guaranteed to be the ad for which the mechanism designer would want to submit the highest such bid. Thus even when there are multiple ads in the same auction with unknown eCPMs, one can continue to make optimal decisions by computing bids for the advertisers equal to their estimated eCPMs plus a value of learning term for the ad and then rank the advertisers on this basis.

10. Performance Guarantees

We now return to the baseline setting in Section 2. The results in the previous sections suggest a possible algorithm that will approximate the optimal bidding strategies for an auctioneer who seeks to maximize long-run efficiency. This algorithm would compute 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 ad 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 ad, where δ is the auctioneer's discount factor, and showing the ad with the highest such score. We refer to this strategy as the *approximately optimal bidding strategy*, and in this section we address questions related to the size of the performance guarantees that can be obtained 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 in the machine learning literature. One other algorithm that is standard for multi-armed bandit problems involves ranking the arms by a term equal to the expected value of the arm plus a term proportional to the standard deviation in the arm (Auer et al., 2002). More generally, one can rank advertisers by a term equal to the eCPM of the advertiser plus a term 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 17 *Suppose the auctioneer uses a bid for the advertiser with unknown eCPM of the form $z = \overline{x} + \frac{c(\overline{x})}{k^\alpha}$, where $\alpha \leq \frac{1}{2}$ and $c(\overline{x})$ is a bounded non-negative constant that depends only on $\overline{x}$ and the distribution of competing bids. Then the optimal constant $c(\overline{x})$ for any such algorithm is $c(\overline{x}) = 0$ for sufficiently large k.*

This result immediately implies that standard existing 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 always making a bid equal to the eCPM of the ad. These existing algorithms do too much exploration, and as a result, lead to lower payoffs than not doing any active exploration at all.[6]

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

Theorem 18 *Suppose the auctioneer follows the approximately optimal bidding strategy. 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 $\frac{\delta^2}{8(1-\delta)^3 k^4} s^4(\overline{x}) f^3(\overline{x}) + o(\frac{1}{k^4})$.*[7]

Theorem 18 indicates that the performance improvement that can be obtained as a result of using the approximately optimal bidding strategy 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 eCPMs is small.

6. Similarly, an algorithm such as epsilon-greedy, in which the ad with the highest eCPM is chosen with probability $1 - \epsilon$, and an ad is chosen uniformly at random with probability ϵ, will also lead to lower payoffs than not doing any active exploration at all for large k. We prove this in Observation 23 in the appendix.
7. The expected payoff increase that we refer to in this theorem is for the subgame beginning from the point when the ad with uncertain eCPM has already received k impressions.

It is worth noting, however, that the result in Theorem 18 is not due to our algorithm being a suboptimal implementation of incorporating active exploration. Our next result illustrates that while the size of the performance improvement that can be obtained by 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 eCPMs.

Theorem 19 *Suppose the auctioneer uses the approximately optimal bidding strategy. 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.*

The results in the previous theorems suggest that the maximum possible payoff increase that can be achieved by incorporating active exploration is quite small for auctions involving ads that have already received a large number of impressions. However, in many auctions, there are frequently 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. Under the mild technical condition discussed in Theorem 13, where the expected value of future learning is lower after the advertiser with unknown eCPM has been shown once rather than never having been shown at all, we obtain the following result:

Theorem 20 *Suppose the bidder with unknown eCPM has a cost-per-click bid of* 1 *and a click-through rate drawn from a beta distribution. 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 supremum of $f(\cdot)$.*

Theorem 20 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 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 20 indicates that the maximum fractional increase in expected payoff that one can achieve from using the theoretically optimal strategy rather than the greedy strategy is roughly $\frac{\delta^2\gamma^8\omega^5\overline{f}^3}{8\xi(1-\delta)^2(1-\omega)^2}$.

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 Lahaie and McAfee (2011) and Ostrovsky and Schwarz (2009) 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 fractional increase in expected payoff that one can achieve from using the theoretically optimal strategy rather than the greedy strategy is roughly $\frac{\delta^2\gamma^8\omega^5\overline{f}^3}{8\xi(1-\delta)^2(1-\omega)^2}$, it then follows that the maximum fractional increase in expected payoff that one can achieve from using the theoretically optimal strategy rather than the greedy strategy is roughly $\frac{\delta^2\gamma^8\omega^2}{8\xi^4(1-\delta)^2(1-\omega)^2}$.

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 $\frac{1}{1000}$ for search ads and display ads respectively (Bax et al., 2011), 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 $\frac{\gamma^8\omega^2}{8\xi^4(1-\omega)^2}$ will almost certainly be less than 10^{-11} in search auctions and 10^{-13} 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), $\frac{\gamma^8\omega^2}{8\xi^4(1-\omega)^2}\frac{\delta^2}{(1-\delta)^2}$ will be no greater than 0.001. Thus as long as $\delta \leq 0.9999$ (or $\delta \leq 0.99999$ for display auctions), the bound given in Theorem 20 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.

11. Simulations

The results of the previous section suggest that the overall benefit that can be obtained by incorporating active exploration 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 CPM bidders in each auction. The CPC bidder has a CPC bid of 1 and a fixed unknown click-through rate. The CPM bidders' CPM bids vary from period to period, and in each period, we assume that the highest CPM bid is a random draw from a distribution with probability density function $f(\cdot)$. Throughout we assume that payoffs are discounted at a rate of $\delta = 0.9995$ and that there are $T = 10000$ time periods.

While we are not aware of any empirical evidence regarding the form of the uncertainty of an advertiser's click-through rate, for simplicity we assume that the CPC bidder's click-through rate is initially drawn from a beta distribution with parameters α and β. The auctioneer may refine this estimate 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 α_t and β_t where α_t is equal to α plus the number of clicks the CPC bidder has received so far and β_t is equal to β plus the number of times the CPC bidder's ad was shown but did not receive a click.

We compare total 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 $\overline{x}_t + \frac{\delta(1-\delta^{T-t})}{2(1-\delta)} \frac{\alpha_t \beta_t}{(\alpha_t+\beta_t)^2(\alpha_t+\beta_t+1)^2} f(\overline{x}_t)$ in each period t, where $\overline{x}_t$ denotes the CPC bidder's expected click-through rate just before the auction in period t. 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 $\frac{1}{1000}$ (Bax et al., 2011), we focus on situations in which the expected click-through rate of the CPC bidder is on the order of $\frac{1}{100}$.

Similarly, since it is unlikely that there will be substantial errors in the estimate of a new ad's predicted click-through rate, we focus on situations in which there is only moderate uncertainty in the click-through rate of a new ad. In particular, 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. We thus consider values of α and β satisfying $(\alpha, \beta) = (10, 1000)$ and $(20, 2000)$ (for 30% and 20% standard errors respectively).

Finally, since there is evidence that the distribution of highest bids is well modeled by a lognormal distribution (Lahaie and McAfee, 2011; Ostrovsky and Schwarz, 2009), we assume throughout that the CPM bidder's bid is drawn from a lognormal distribution with parameters μ and σ^2. We use a value of $\sigma^2 = \log(2)$ to match the variance in the lognormal distribution estimated by Ostrovsky and Schwarz (2009). And Varian (2009) has noted that the total value enjoyed by advertisers is typically about $2 - 2.3$ times their total expenditure. If the auction consisted of only two advertisers, this would suggest that the appropriate value of μ would be such that the highest competing bidder had a CPM bid that is roughly double that of the CPC bidder in expectation. However, since there are more than two bidders in most real auctions, the appropriate value of μ will be larger than this. We thus consider a range of values of μ from -4.25 (for the case in which the highest competing CPM bid is roughly double that of the CPC bidder in expectation) to -3.5 (for the case in which the highest competing CPM bid is roughly four times that of the CPC bidder in expectation).

Table 1 reports the results our simulations. 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 in Table 1 are statistically significant. Indeed one can conclude from these simulations that the maximum possible efficiency gain

Conditions	Percentage increase in efficiency
$\alpha = 10, \beta = 1000, \mu = -4.25, \sigma^2 = \log(2)$	0.021% (0.017%)
$\alpha = 10, \beta = 1000, \mu = -4, \sigma^2 = \log(2)$	-0.016% (0.011%)
$\alpha = 10, \beta = 1000, \mu = -3.75, \sigma^2 = \log(2)$	-0.008% (0.007%)
$\alpha = 10, \beta = 1000, \mu = -3.5, \sigma^2 = \log(2)$	0.003% (0.004%)
$\alpha = 20, \beta = 2000, \mu = -4.25, \sigma^2 = \log(2)$	0.003% (0.009%)
$\alpha = 20, \beta = 2000, \mu = -4, \sigma^2 = \log(2)$	0.001% (0.006%)
$\alpha = 20, \beta = 2000, \mu = -3.75, \sigma^2 = \log(2)$	0.001% (0.004%)
$\alpha = 20, \beta = 2000, \mu = -3.5, \sigma^2 = \log(2)$	-0.002% (0.002%)

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

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 in an auction setting is exceedingly small.

The reason for the results observed in Table 1 is that an optimal exploration algorithm will only do a tiny additional amount of exploration compared to the greedy strategy of always submitting a bid for the CPC bidder equal to the CPC bidder's estimated eCPM. 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 4.2%, implying only a 1.8% increase in the probability that the CPC bidder will be shown as well as only a 2.1% 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 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 in 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 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 click-through rate 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.

Conditions	Percentage increase in efficiency
$\alpha = 2, \beta = 200, \mu = -4, \sigma^2 = \log(2)/4$	0.15% (0.05%)
$\alpha = 2, \beta = 200, \mu = -3.75, \sigma^2 = \log(2)/4$	0.17% (0.05%)

Table 2: Average percentage increase in efficiency from incorporating active learning (with standard errors in parentheses) after 10000 simulations. These results are both statistically significant at the $p < .005$ 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.

Finally, while the gains achieved through active exploration in Table 2 are small, one would not achieve greater gains by using a standard algorithm such as UCB. To test this, we considered the same setting in the first row of this table, but instead of making a bid for the CPC bidder of the form $\overline{x}_t + \frac{\delta(1-\delta^{T-t})}{2(1-\delta)} \frac{\alpha_t \beta_t}{(\alpha_t+\beta_t)^2(\alpha_t+\beta_t+1)^2} f(\overline{x}_t)$ in each period t, we made a bid of the form $\overline{x}_t + c(\overline{x}_t) \frac{1}{\sqrt{\alpha_t+\beta_t}}$, where the constant $c(\overline{x}_t)$ was chosen so that this bid would equal $\overline{x}_t + \frac{\delta(1-\delta^{T-t})}{2(1-\delta)} \frac{\alpha_t \beta_t}{(\alpha_t+\beta_t)^2(\alpha_t+\beta_t+1)^2} f(\overline{x}_t)$ in time period $t = 1$. Thus our implementation of UCB performed the same amount of exploration as the main algorithm we considered in the very first period of the game, while performing more exploration in later periods due to the fact that the rate of exploration declines with $\frac{1}{(\alpha_t+\beta_t)^2}$ under our proposed algorithm, while only declining with $\frac{1}{\sqrt{\alpha_t+\beta_t}}$ under UCB.

In this setting, we found that using the UCB algorithm rather than the purely greedy strategy resulted in an average efficiency loss of 1.04% (with a standard error of 0.07%). Thus while we were able to achieve an improvement by using the new algorithm we have proposed, using the UCB algorithm instead resulted in significant efficiency losses. The fact that UCB performed worse than the purely greedy strategy is not surprising since we know from Theorem 17 that UCB performs worse than the purely greedy strategy once an ad has received enough impressions.

12. Conclusion

In online auctions, there may be value to exploring ads with uncertain eCPMs 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 because there may be considerable 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 is dramatically different than the optimal solution in standard multi-armed bandit problems, and in particular, that the optimal amount of active exploration is considerably smaller than in standard multi-armed bandit problems. This in turn implies that the improvement in the auctioneer's payoff that can be achieved by adding active learning in online auctions is also exceedingly small. Thus while it is theoretically possible to improve efficiency by incorporating active learning, in a practical exchange environment, a purely greedy strategy of simply ranking the ads by their expected eCPMs is likely to perform nearly as well as any other strategy.

We conclude by discussing one other point. Throughout our analysis we have focused on the problem of an auctioneer who wants to maximize efficiency. Although this is a sensible objective, one might also envision scenarios in which the mechanism designer wishes to maximize a weighted average of efficiency and revenue. While incorporating active exploration in online auctions can only have a small effect on efficiency, this active exploration may significantly improve revenue. The reason for this is that if we rank the ads by the sum of their expected eCPMs and a value of learning term, the value of learning term may be larger for ads that typically lose the auctions, and incorporating this value of learning term may increase pricing pressure for the winning ads and thereby increase revenue.[8] In fact, in several of the simulations considered in the previous section in which incorporating active exploration failed to show significant efficiency gains, the algorithm that we considered still showed significant revenue gains over the purely greedy strategy of ranking the ads by their expected eCPMs. But while it is still possible to achieve significant revenue gains by incorporating active exploration in the type of environment considered in this paper, the maximum possible efficiency gains are likely to be exceedingly small.

Acknowledgments

We especially thank Martin Zinkevich for numerous helpful discussions. We are also grateful to Joshua Dillon, Pierre Grinspan, Chris Harris, Tim Lipus, Mohammad Mahdian, Hal Varian, and the anonymous referees for helpful comments and discussions.

This work is based on an earlier work: "Machine Learning in an Auction Environment", in *Proceedings of the* 23^{rd} *International Conference on the World Wide Web* (WWW) (2014) http://doi.acm.org/10.1145/2566486.2567974.

8. A similar point has been previously noted by Li et al. (2010) and McAfee (2011).

Appendix A. Proofs of Theorems

Proof of Theorem 1: Suppose it is known that the eCPM of the ad is x. If the highest eCPM for a competing ad is p, then the presence of this ad with eCPM x increases total welfare by $x - p$ if $x > p$ and 0 otherwise. Thus the expected increase in total welfare from this ad with eCPM of x competing in the auction is $\int_0^x (x-p)f(p)\ dp$. The total long-term value from having this advertisement is then the discounted sum of this expected total increase in welfare or $\frac{1}{1-\delta}\int_0^x (x-p)f(p)\ dp$.

Now if $V(x) \equiv \frac{1}{1-\delta}\int_0^x (x-p)f(p)\ dp$, then $V'(x) = \frac{1}{1-\delta}\int_0^x f(p)\ dp$ and $V''(x) = \frac{1}{1-\delta}f(x)$. From this it follows that $V''(x) \geq 0$ for all x and $V''(x) > 0$ if x is contained in the support of F. Thus the long-term value of the advertisement is a convex function of the eCPM of the ad and a strictly convex function if the eCPM of the ad is contained within the support of the distribution of the highest competing eCPM. ■

Proof of Lemma 2: Suppose an ad has been shown k times. The value of the dynamic program that arises from placing the optimal bid z in the current period, $V_k(\overline{x})$, equals the immediate reward from bidding z (or the negative of the loss function) in the current period plus δ times the expected value of the dynamic program 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 program that arises next period is $E_{\overline{x}'}[V_{k+1}(\overline{x}')]$, where the expectation is taken over the randomness in the changes in the estimates of the eCPM of the ad $\overline{x}'$ 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 program remains at $V_k(\overline{x})$. Thus the expected value of the dynamic program that arises in the next period is $F(z)E_{\overline{x}'}[V_{k+1}(\overline{x}')] + (1 - F(z))V_k(\overline{x})$.

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

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

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

$$V_k(\overline{x}) = \frac{1}{1-\delta}\left(\max_z E_\epsilon \left[-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy + \delta F(z)(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})) \right] \right). \blacksquare$$

Proof of Theorem 3: By differentiating the expression in Lemma 2 with respect to z, we see that the first order condition for z to be an optimal bid is

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

From this it follows that $z = \overline{x} + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$ satisfies the first order conditions. Moreover, at this value of z, the second order conditions are also satisfied. Thus optimal bidding entails setting $z = \overline{x} + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$. ■

Proof of Lemma 4: Let $\Phi(\cdot|\sigma_k)$ denote the distribution from which ϵ is drawn for any given value of σ_k. For any given σ_k, we know that $\Phi(\cdot|\sigma_k)$ has mean zero and variance one. We also know from the Bayesian central limit theorem that as $\sigma_k \to 0$ (and $k \to \infty$) that $\Phi(\cdot|\sigma_k)$ converges to the standard normal distribution. For any given σ_k, we can write $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy]$ as $J(\sigma_k) = E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)]$. We seek to show that $J(\sigma_k)$ is of the form given in the statement of the lemma.

First note that

$$J'(\sigma_k) = -E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\sigma_k)] + \frac{d}{d\Phi}E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)\right]$$

where $\frac{d}{d\Phi}E_\epsilon[Z(\epsilon,\sigma_k)|\epsilon \sim \Phi(\cdot|\sigma_k)]$ denotes the derivative of the expectation of $Z(\epsilon,\sigma_k)$ arising through the changes in $\Phi(\cdot|\sigma_k)$ induced by changes in σ_k (that is, if $\phi(\epsilon;\sigma_k)$ denotes the density corresponding to $\Phi(\cdot|\sigma_k)$, then $\frac{d}{d\Phi}E_\epsilon[Z(\epsilon,\sigma_k)|\epsilon \sim \Phi(\cdot|\sigma_k)] \equiv \int_{-\infty}^{\infty} Z(\epsilon,\sigma_k)\frac{\partial\phi}{\partial\sigma_k}(\epsilon;\sigma_k)d\epsilon)$. Similarly, letting $\frac{d^m}{d\Phi^m}E_\epsilon[Z(\epsilon,\sigma_k)|\epsilon \sim \Phi(\cdot|\sigma_k)] \equiv \int_{-\infty}^{\infty} Z(\epsilon,\sigma_k)\frac{\partial^m\phi}{\partial\sigma_k^m}(\epsilon;\sigma_k)d\epsilon$ for all m, we have

$$\begin{aligned} J''(\sigma_k) &= E_\epsilon[\epsilon^2 f(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] - 2\frac{d}{d\Phi}E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad + \frac{d^2}{d\Phi^2}E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)\right], \end{aligned}$$

$$\begin{aligned} J'''(\sigma_k) &= E_\epsilon[\epsilon^3 f'(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] + 3\frac{d}{d\Phi}E_\epsilon[\epsilon^2 f(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad - 3\frac{d^2}{d\Phi^2}E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad + \frac{d^3}{d\Phi^3}E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)\right], \end{aligned}$$

and

$$\begin{aligned} J''''(\sigma_k) &= E_\epsilon[\epsilon^4 f''(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] + 4\frac{d}{d\Phi}E_\epsilon[\epsilon^3 f'(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad + 6\frac{d^2}{d\Phi^2}E_\epsilon[\epsilon^2 f(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad - 4\frac{d^3}{d\Phi^3}E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\cdot|\sigma_k)] \\ &\quad + \frac{d^4}{d\Phi^4}E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)\right], \end{aligned}$$

Note that when $\sigma_k = 0$, we have $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)] = \int_{\overline{x}}^{z} F(z) - F(y)\ dy$ for any distribution $\Phi(\cdot|\sigma_k)$, $E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\cdot|\sigma_k)] = E_\epsilon[\epsilon(F(z) - F(\overline{x}))|\epsilon \sim \Phi(\cdot|\sigma_k)] = 0$ for any distribution $\Phi(\cdot|\sigma_k)$ with mean zero, and $E_\epsilon[\epsilon^2 f(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] = E_\epsilon[\epsilon^2 f(\overline{x})|\epsilon \sim \Phi(\cdot|\sigma_k)] = f(\overline{x})$ for any distribution $\Phi(\cdot|\sigma_k)$ with mean zero and variance one. Thus $\frac{d^m}{d\Phi^m}E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy|\epsilon \sim \Phi(\cdot|\sigma_k)] = 0$, $\frac{d^m}{d\Phi^m}E_\epsilon[\epsilon(F(z) - F(\overline{x}+\sigma_k\epsilon))|\epsilon \sim \Phi(\cdot|\sigma_k)] = 0$, and $\frac{d^m}{d\Phi^m}E_\epsilon[\epsilon^2 f(\overline{x}+\sigma_k\epsilon)|\epsilon \sim \Phi(\cdot|\sigma_k)] = 0$ for all m when evaluated at $\sigma_k = 0$.

By using these facts, the fact that $\Phi(\cdot|0)$ is standard normal, and the above expressions for $J(\sigma_k)$ and its derivatives, it follows that $J(0) = \int_{\overline{x}}^{z} F(z) - F(y)\ dy$, $J'(0) = 0$, $J''(0) = f(\overline{x})$, $J'''(0) = 0$, and $J''''(0) = E_\epsilon[\epsilon^4 f''(\overline{x})|\epsilon \sim \Phi(\cdot|0)] + 4\frac{d}{d\Phi}E_\epsilon[\epsilon^3 f'(\overline{x})|\epsilon \sim \Phi(\cdot|\sigma_k)]|_{\sigma_k=0}$. This in turn implies that the fourth-order Taylor approximation to $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z)-F(y)\ dy]$ is

$$E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy\right] = \int_{\overline{x}}^{z} F(z) - F(y)\ dy + \frac{1}{2}\sigma_k^2 f(\overline{x}) + a(\overline{x})\sigma_k^4 + o(\sigma_k^4),$$

where $a(\overline{x}) \equiv \frac{1}{24}[E_\epsilon[\epsilon^4 f''(\overline{x})|\epsilon \sim \Phi(\cdot|0)] + 4\frac{d}{d\Phi}E_\epsilon[\epsilon^3 f'(\overline{x})|\epsilon \sim \Phi(\cdot|\sigma_k)]|_{\sigma_k=0}]$. ■

Proof of Theorems 5 and 6: Since these results are special cases of Theorems 9 and 10 respectively, the proofs of these results are omitted.

Before proving Theorem 7, we first introduce some notation for the finite-horizon version of this game. If the game has a finite time horizon and will last an additional T periods, we let $V_{k,T}(\overline{x})$ denote the value of the dynamic program that arises when the auctioneer follows the optimal strategy. By analogy to Lemma 2, we know that

$$V_{k,T}(\overline{x}) = \frac{1}{1-\delta}\left(\max_z E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy + \delta F(z)(E_{\overline{x}'}[V_{k+1,T-1}(\overline{x}')] - V_{k,T-1}(\overline{x}))\right]\right)$$

when $T > 0$ and $V_{k,T}(\overline{x}) = -E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{\overline{x}} F(\overline{x}) - F(y)\ dy]$ when $T = 0$. Also note that $\lim_{T\to\infty} V_{k,T}(\overline{x}) = V_k(\overline{x})$, where $V_k(\overline{x})$ is the value of the dynamic program for the original infinite-horizon game. Finally note that

Lemma 21 *$V_{k,T}(\overline{x})$ is twice differentiable in $\overline{x}$ for all k and T. Furthermore, $\lim_{k\to\infty} V'_{k,T}(\overline{x}) = 0$ and $\lim_{k\to\infty} V''_{k,T}(\overline{x}) = 0$ for all T.*

Proof We prove this result by induction on T. The base case, $T = 0$, holds because the fact that $f(\cdot)$ is continuously differentiable implies $F(\cdot)$ is twice differentiable and $V_{k,T}(\overline{x}) = -E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{\overline{x}} F(\overline{x}) - F(y)\ dy]$ is also twice differentiable in $\overline{x}$. Furthermore, $V'_{k,T}(\overline{x}) = E_\epsilon[F(\overline{x}) - F(\overline{x}+\sigma_k\epsilon) - \int_{\overline{x}+\sigma_k\epsilon}^{\overline{x}} f(\overline{x})\ dy] = E_\epsilon[F(\overline{x}) - F(\overline{x}+\sigma_k\epsilon)]$ and $V''_{k,T}(\overline{x}) = E_\epsilon[f(\overline{x}) - f(\overline{x}+\sigma_k\epsilon)]$, which both tend to zero as $k \to \infty$. Thus the result holds for $T = 0$.

Now suppose the result is true for $T-1$ and use this to prove the result must also hold for T. Since

$$V_{k,T}(\overline{x}) = \frac{1}{1-\delta}\left(\max_z E_\epsilon\left[-\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy + \delta F(z)(E_{\overline{x}'}[V_{k+1,T-1}(\overline{x}'))] - V_{k,T-1}(\overline{x}))\right]\right),$$

we know from analogy to Theorem 3 that the optimal bid z satisfies $z = \overline{x}+\delta(E_{\overline{x}'}[V_{k+1,T-1}(\overline{x}'))]-V_{k,T-1}(\overline{x}))$. From the induction hypothesis, we thus know that the optimal bid $z_k(\overline{x})$ is twice differentiable in $\overline{x}$ and that $\lim_{k\to\infty} z'_k(\overline{x}) = 1$ and $\lim_{k\to\infty} z''_k(\overline{x}) = 0$.

This in turn implies that $E_\epsilon[-\int_{\overline{x}+\sigma_k\epsilon}^{z_k(\overline{x})} F(z_k(\overline{x})) - F(y)\ dy]$ is twice differentiable in $\overline{x}$ and that $\frac{d}{d\overline{x}}E_\epsilon[-\int_{\overline{x}+\sigma_k\epsilon}^{z_k(\overline{x})} F(z_k(\overline{x}))-F(y)\ dy] = E_\epsilon[F(z_k(\overline{x}))-F(\overline{x}+\sigma_k\epsilon)-\int_{\overline{x}+\sigma_k\epsilon}^{z_k(\overline{x})} z'_k(\overline{x})f(z_k(\overline{x}))\ dy]$, which tends to zero as $k \to \infty$ since $\lim_{k\to\infty} z_k(\overline{x}) = \overline{x}$. This also further implies that $\frac{d^2}{d^2\overline{x}}E_\epsilon[-\int_{\overline{x}+\sigma_k\epsilon}^{z_k(\overline{x})} F(z_k(\overline{x}))-F(y)\ dy] = E_\epsilon[z'_k(\overline{x})f(z_k(\overline{x}))-f(\overline{x}+\sigma_k\epsilon)-(z'_k(\overline{x})-1)z'_k(\overline{x})f(z_k(\overline{x}))-$

$\int_{\overline{x}+\sigma_k\epsilon}^{z_k(\overline{x})} z_k''(\overline{x})f(z_k(\overline{x}))+(z_k'(\overline{x}))^2 f'(z_k(\overline{x}))\,dy]$, which tends to zero as $k \to \infty$ since $\lim_{k\to\infty} z_k(\overline{x}) = \overline{x}$ and $\lim_{k\to\infty} z_k'(\overline{x}) = 1$.

From the induction hypothesis, we also know that $F(z_k(\overline{x}))(E_{\overline{x}'}[V_{k+1,T-1}(\overline{x}')]-V_{k,T-1}(\overline{x}))$ is twice differentiable in $\overline{x}$ and that the first and second derivatives of this expression with respect to $\overline{x}$ tend to zero as $k \to \infty$. By combining this with the results in the previous two paragraphs, it follows that $V_{k,T}(\overline{x})$ is twice differentiable in $\overline{x}$ and $\lim_{k\to\infty} V_{k,T}'(\overline{x}) = 0$ and $\lim_{k\to\infty} V_{k,T}''(\overline{x}) = 0$ for all T. The result follows by induction. ■

We use these observations about the finite-horizon game to first prove that $E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = O(\frac{1}{k^2})$ in Lemma 22. Then we use this preliminary result to prove Theorem 7.

Lemma 22 *$E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = O(\frac{1}{k^2})$ 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 $\overline{x}_c$ denote the estimated eCPM of the ad if the ad receives a click, and let $\overline{x}_n$ denote the estimated eCPM of the ad if the ad does not receive a click. Note that $p\overline{x}_c + (1-p)\overline{x}_n = \overline{x}$.

From Lemma 21 we know that $V_{k,T}(\overline{x})$ is twice differentiable in $\overline{x}$ for all k and T. Thus the second-order Taylor approximations for $V_{k+1,T}(\overline{x}_c)$ and $V_{k+1,T}(\overline{x}_n)$ are

$$V_{k+1,T}(\overline{x}_c) = V_{k+1,T}(\overline{x}) + V_{k+1,T}'(\overline{x})(\overline{x}_c - \overline{x}) + \frac{1}{2}V_{k+1,T}''(\overline{x})(\overline{x}_c - \overline{x})^2 + o(\overline{x}_c - \overline{x})^2$$

and

$$V_{k+1,T}(\overline{x}_n) = V_{k+1,T}(\overline{x}) + V_{k+1,T}'(\overline{x})(\overline{x}_n - \overline{x}) + \frac{1}{2}V_{k+1,T}''(\overline{x})(\overline{x}_n - \overline{x})^2 + o(\overline{x}_n - \overline{x})^2.$$

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 using 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

$$\begin{aligned} E[V_{k+1,T}(\overline{x}')] &= pV_{k+1,T}(\overline{x}_c) + (1-p)V_{k+1,T}(\overline{x}_n) \\ &= V_{k+1,T}(\overline{x}) + \frac{1}{2}V_{k+1,T}''(\overline{x})E[(\overline{x}' - \overline{x})^2] + o(E[(\overline{x}' - \overline{x})^2]). \end{aligned}$$

From this it follows that

$$E[V_{k+1,T}(\overline{x}') - V_{k,T}(\overline{x})] = V_{k+1,T}(\overline{x}) - V_{k,T}(\overline{x}) + \frac{1}{2}V_{k+1,T}''(\overline{x})E[(\overline{x}' - \overline{x})^2] + o(E[(\overline{x}' - \overline{x})^2]). \quad (1)$$

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 $(\overline{x}' - \overline{x})^2 = O(\frac{1}{k^2})$. Furthermore, from Lemma 21 we know that $\lim_{k\to\infty} V''_{k+1,T}(\overline{x}) = 0$. Thus we can rewrite equation (1) as

$$E[V_{k+1,T}(\overline{x}') - V_{k,T}(\overline{x})] = V_{k+1,T}(\overline{x}) - V_{k,T}(\overline{x}) + o\left(\frac{1}{k^2}\right).$$

By using the fact that $\lim_{T\to\infty} V_{k,T}(\overline{x}) = V_k(\overline{x})$, where $V_k(\overline{x})$ denotes the value of the dynamic program in the original infinite horizon game, we then know that

$$\begin{aligned} E[V_{k+1}(\overline{x}') - V_k(\overline{x})] &= V_{k+1}(\overline{x}) - V_k(\overline{x}) + o\left(\frac{1}{k^2}\right) \\ &= \frac{v(\overline{x})}{k} - \frac{v(\overline{x})}{k+1} + O\left(\frac{1}{k^2}\right) \\ &= O\left(\frac{1}{k^2}\right). \end{aligned}$$

■

Proof of Theorem 7: We have seen in the proof of Lemma 22 that $E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = V_{k+1}(\overline{x}) - V_k(\overline{x}) + o\left(\frac{1}{k^2}\right)$. When combined with the fact that $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + O(\frac{1}{k^2})$, this immediately implied that $E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = O\left(\frac{1}{k^2}\right)$. If we are able to further prove that we can write $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + \frac{w(\overline{x})}{k^2} + o(\frac{1}{k^2})$ for some function $w(\overline{x})$, it will then follow that $V_{k+1}(\overline{x}) - V_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o(\frac{1}{k^2})$. Thus we first seek to show that we can write $V_k(\overline{x})$ as $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + \frac{w(\overline{x})}{k^2} + o(\frac{1}{k^2})$.

Since $E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = O(\frac{1}{k^2})$ for large k and the optimal bidding strategy entails setting $z = \overline{x} + \delta(E[V_{k+1}(\overline{x}') - V_k(\overline{x})])$, it must be the case that $z = \overline{x} + O(\frac{1}{k^2})$ for large k. From this it follows that $\int_{\overline{x}}^{z} F(z) - F(y)\ dy = o(\frac{1}{k^2})$ under the optimal bidding strategy z for large k.

Now we have seen in Lemma 4 that $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \int_{\overline{x}}^{z} F(z) - F(y)\ dy + \frac{1}{2}\sigma_k^2 f(\overline{x}) + a(\overline{x})\sigma_k^4 + o(\sigma_k^4)$ for some constant $a(\overline{x})$ for large k. Since $\int_{\overline{x}}^{z} F(z) - F(y)\ dy = o(\frac{1}{k^2})$ under the optimal bidding strategy z and $\sigma_k^2 = \frac{s^2(\overline{x})}{k} + \frac{h(\overline{x})}{k^2} + o(\frac{1}{k^2})$ for large k, it then follows that $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \frac{1}{2k}s^2(\overline{x})f(\overline{x}) + \frac{1}{k^2}[\frac{h(\overline{x})f(\overline{x})}{2} + a(\overline{x})s^4(\overline{x})] + o(\frac{1}{k^2})$ for large k, which we can rewrite as $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \frac{1}{2k}s^2(\overline{x})f(\overline{x}) + \frac{1}{k^2}u(\overline{x}) + o(\frac{1}{k^2})$, where $u(\overline{x}) \equiv \frac{h(\overline{x})f(\overline{x})}{2} + a(\overline{x})s^4(\overline{x})$.

But $-E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy]$ represents the auctioneer's per-period payoff in the next auction. Thus if $\overline{x}'$ denotes the estimated eCPM of the ad after an additional j periods have passed and k' denotes the number of impressions the ad has received after an additional j periods have passed, then the auctioneer's per-period payoff in the period after an additional j periods have passed is $-\frac{1}{2k'}s^2(\overline{x}')f(\overline{x}') - \frac{1}{2k'^2}u(\overline{x}') + o(\frac{1}{k'^2})$. The difference between this and $-\frac{1}{2k}s^2(\overline{x})f(\overline{x})$ is

$$-\frac{1}{2k'}s^2(\overline{x}')f(\overline{x}') - \frac{1}{2k'^2}u(\overline{x}') + \frac{1}{2k}s^2(\overline{x})f(\overline{x}) + o\left(\frac{1}{k^2}\right)$$

$$
\begin{aligned}
&= \frac{k's^2(\overline{x})f(\overline{x}) - ks^2(\overline{x}')f(\overline{x}')}{2kk'} - \frac{1}{2k^2}u(\overline{x}) + \left[\frac{1}{2k^2}u(\overline{x}) - \frac{1}{2k'^2}u(\overline{x}')\right] + o\left(\frac{1}{k^2}\right) \\
&= \frac{k's^2(\overline{x})f(\overline{x}) - ks^2(\overline{x}')f(\overline{x}')}{2kk'} - \frac{1}{2k^2}u(\overline{x}) + \frac{k'^2u(\overline{x}) - k^2u(\overline{x}')}{2k^2k'^2} + o\left(\frac{1}{k^2}\right) \\
&= \frac{k's^2(\overline{x})f(\overline{x}) - k[s^2(\overline{x})f(\overline{x}) + (\overline{x}' - \overline{x})d(\overline{x}) + o(\overline{x}' - \overline{x})]}{2kk'} - \frac{1}{2k^2}u(\overline{x}) \\
&\quad + \frac{k'^2u(\overline{x}) - k^2[u(\overline{x}) + O(\overline{x}' - \overline{x})]}{2k^2k'^2} + o\left(\frac{1}{k^2}\right), \qquad (2)
\end{aligned}
$$

where $d(\overline{x})$ denotes the derivative of the function $s^2(\overline{x})f(\overline{x})$ with respect to $\overline{x}$. By the same reasoning as in the proof of Lemma 22, we know that $\overline{x}' - \overline{x} = O(\frac{1}{k})$ for all possible realizations of $\overline{x}'$. Thus we can rewrite the expression in equation (2) as

$$
\begin{aligned}
&\frac{(k' - k)s^2(\overline{x})f(\overline{x}) - k(\overline{x}' - \overline{x})d(\overline{x})}{2kk'} - \frac{1}{2k^2}u(\overline{x}) + \frac{(k'^2 - k^2)u(\overline{x})}{2k^2k'^2} + o\left(\frac{1}{k^2}\right) \\
= &\frac{(k' - k)s^2(\overline{x})f(\overline{x}) - k(\overline{x}' - \overline{x})d(\overline{x})}{2kk'} - \frac{1}{2k^2}u(\overline{x}) + o\left(\frac{1}{k^2}\right). \qquad (3)
\end{aligned}
$$

Now note that $E[\overline{x}'] = \overline{x}$, where the expectation is taken over the uncertain realization of $\overline{x}'$ in another j periods. Thus the expectation of the expression in equation (3) is

$$
E\left[\frac{(k' - k)s^2(\overline{x})f(\overline{x})}{2kk'}\right] - \frac{1}{2k^2}u(\overline{x}) + o\left(\frac{1}{k^2}\right), \qquad (4)
$$

where the expectation is taken over the uncertain realization of k'. This expression can in turn be written as

$$
\frac{\overline{m}_j(\overline{x})s^2(\overline{x})f(\overline{x}) - u(\overline{x})}{2k^2} + o\left(\frac{1}{k^2}\right), \qquad (5)
$$

where $\overline{m}_j(\overline{x})$ denotes the expected number of additional impressions that the ad with uncertain eCPM receives after an additional j periods have passed (which will equal 0 when $j = 0$ and vary approximately linearly with j for large k).

The expression in equation (5) gives the difference between the auctioneer's actual expected payoff in the period after an additional j periods have passed and $-\frac{1}{2k}s^2(\overline{x})f(\overline{x})$. From this it follows that the difference between the auctioneer's actual payoff $V_k(\overline{x})$ and the payoff the auctioneer would receive if the auctioneer obtained a payoff of $-\frac{1}{2k}s^2(\overline{x})f(\overline{x})$ in every future period is $\sum_{j=0}^{\infty}\delta^j\frac{\overline{m}_j(\overline{x})s^2(\overline{x})f(\overline{x}) - u(\overline{x})}{2k^2} + o\left(\frac{1}{k^2}\right)$, which can be written as $\frac{w(\overline{x})}{k^2} + o(\frac{1}{k^2})$ for some function $w(\overline{x})$. Thus we can write $V_k(\overline{x})$ as $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + \frac{w(\overline{x})}{k^2} + o(\frac{1}{k^2})$ for some function $w(\overline{x})$.

But we have seen in the proof of Lemma 22 that $E[V_{k+1}(\overline{x}') - V_k(\overline{x})] = V_{k+1}(\overline{x}) - V_k(\overline{x}) + o\left(\frac{1}{k^2}\right)$. Since $V_k(\overline{x}) = -\frac{v(\overline{x})}{k} + \frac{w(\overline{x})}{k^2} + o(\frac{1}{k^2})$, it then follows that $V_{k+1}(\overline{x}) - V_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o(\frac{1}{k^2})$. ■

Proof of Theorem 8: Recall that

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

and a second-order Taylor approximation for $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy]$ is

$$E_\epsilon\left[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy\right] = \int_{\overline{x}}^{z} F(z) - F(y)\ dy + \frac{1}{2}\sigma_k^2 f(\overline{x}) + o(\sigma_k^2)$$

Now $z = \overline{x} + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$, so $z - \overline{x} \leq -\delta V_k(\overline{x})$, a term which is $O(f(\overline{x})\sigma_k^2)$. From this it follows that $\int_{\overline{x}}^{z} F(z) - F(y)\ dy = O(F(\overline{x})f(\overline{x})\sigma_k^2) = o(f(\overline{x})\sigma_k^2)$. And we also know that $\delta F(z)(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})) \leq -\delta F(z)V_k(\overline{x}) = O(F(\overline{x})f(\overline{x})\sigma_k^2) = o(f(\overline{x})\sigma_k^2)$.

Combining these results gives $E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \frac{1}{2}\sigma_k^2 f(\overline{x}) + o(f(\overline{x})\sigma_k^2)$ and $F(z)(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})) = o(f(\overline{x})\sigma_k^2)$. Substituting this in to our expression for $V_k(\overline{x})$ then gives $V_k(\overline{x}) = -\frac{1}{2(1-\delta)} f(\overline{x})\sigma_k^2 + o(f(\overline{x})\sigma_k^2)$. ■

Proof of Theorem 9: First note that it must be the case that $V_{\vec{k}}(\overline{x}) = \Omega(\frac{1}{k})$ for large k. We know that $\sigma_{a,k_a}^2 = \Theta(\frac{1}{k_a}) = \Theta(\frac{1}{\beta_a k}) = \Theta(\frac{1}{k})$ for large k, so the immediate reward in any given period is at least on the same order as $\frac{1}{k}$ regardless of which ad-context pair a arises in the auction. Thus we know that $V_{\vec{k}}(\overline{x}) = \Omega(\frac{1}{k})$ for large k.

We also know that $V_{\vec{k}}(\overline{x}) = O(\frac{1}{k})$ for large k. To see this, note that the auctioneer can ensure that his loss in an auction involving the ad-context pair a in any given period is $O(\frac{1}{k_a}) = O(\frac{1}{k})$ by bidding $z_a = \overline{x}_a$, so the auctioneer can thus ensure that his expected loss in any given period is $O(\frac{1}{k})$ unconditional on the precise ad-context pair that arises. And if the auctioneer's loss in any given period is $O(\frac{1}{k})$, then the player'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 $\Theta(\frac{1}{k})$, $\sum_{j=k}^{\infty} \delta^{j-k}\frac{1}{j}$, 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 $V_{\vec{k}}(\overline{x}) = \Theta(\frac{1}{k})$ for large k. ■

Proof of Theorem 10: Since $V_{\vec{k}}(\overline{x}) = \Theta(\frac{1}{k})$ for large k, we have $E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x}) = O(\frac{1}{k})$ for large k. Thus since the optimal bidding strategy entails setting $z_a = \overline{x}_a + \delta(E_{\overline{x}'(a)}[V_{\vec{k}'(a,\vec{k})}(\overline{x}'(a))] - V_{\vec{k}}(\overline{x}))$, it must be the case that $z_a = \overline{x}_a + O(\frac{1}{k})$ for large k. From this it follows that $\int_{\overline{x}_a}^{z_a} F_a(z_a) - F_a(y)\ dy = O(\frac{1}{k^2})$ under the optimal bidding strategy z_a for large k.

Now we have seen in Lemma 4 that $E_\epsilon[\int_{\overline{x}_a+\sigma_{a,k_a}\epsilon}^{z_a} F_a(z_a) - F_a(y)\ dy] = \int_{\overline{x}_a}^{z_a} F_a(z_a) - F_a(y)\ dy + \frac{1}{2}\sigma_{a,k_a}^2 f_a(\overline{x}_a) + O(\sigma_{a,k_a}^4)$ for large k. Since $\int_{\overline{x}_a}^{z_a} F_a(z_a) - F_a(y)\ dy = O(\frac{1}{k^2})$ under the optimal bidding strategy z_a and $\sigma_{a,k_a}^2 = \frac{s_a^2(\overline{x}_a)}{k_a} + O(\frac{1}{k^2})$ for large k, it then follows that $E_\epsilon[\int_{\overline{x}_a+\sigma_{a,k_a}\epsilon}^{z_a} F_a(z_a) - F_a(y)\ dy] = \frac{1}{2k_a} s_a^2(\overline{x}_a) f_a(\overline{x}_a) + O(\frac{1}{k^2})$ for large k.

But $-E_\epsilon[\int_{\overline{x}_a+\sigma_{a,k_a}\epsilon}^{z_a} F_a(z_a) - F_a(y)\ dy]$ represents the auctioneer's per-period payoff if the next auction is an auction for the advertiser-context pair a. Thus the auctioneer's expected per-period payoff unconditional on what ad-context pair appears in the next auction is $\sum_{a=1}^{m} \pi_a \frac{1}{2k_a} s_a^2(\overline{x}_a) f_a(\overline{x}_a) + O(\frac{1}{k^2}) = \sum_{a=1}^{m} \pi_a \frac{1}{2\beta_a k} s_a^2(\overline{x}_a) f_a(\overline{x}_a) + O(\frac{1}{k^2})$ for large k. From this it follows that if $g(\overline{x}) \equiv \sum_{a=1}^{m} \pi_a \frac{1}{2\beta_a} s_a^2(\overline{x}_a) f_a(\overline{x}_a)$, then the expected per-period utility that one obtains at each point in the game unconditional on what ad-context pair appears in the next auction is $\frac{1}{k} g(\overline{x}) + O(\frac{1}{k^2})$.

Since $V_{\vec{k}}(\overline{x})$ 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 $|kV_{\vec{k}}(\overline{x})| \leq \sum_{j=k}^{\infty} \delta^{j-k}[g(\overline{x})] + O(\frac{1}{k})$, meaning $|kV_{\vec{k}}(\overline{x})| \leq \frac{1}{1-\delta} g(\overline{x}) + O(\frac{1}{k})$ and $|kV_{\vec{k}}(\overline{x})| \geq \sum_{j=k}^{\infty} \delta^{j-k}[\frac{k}{j} g(\overline{x})] +$

$O(\frac{1}{k}) = \frac{1}{1-\delta}g(\overline{x}) + O(\frac{1}{k})$ in the limit as $k \to \infty$. From this it follows that $|kV_{\vec{k}}(\overline{x})| = \frac{1}{1-\delta}g(\overline{x}) + O(\frac{1}{k})$ and $kV_{\vec{k}}(\overline{x}) = -\frac{1}{1-\delta}g(\overline{x}) + O(\frac{1}{k}) = -\frac{1}{2(1-\delta)}\sum_{a=1}^{m}\pi_a\frac{1}{\beta_a}s_a^2(\overline{x}_a)f_a(\overline{x}_a) + O(\frac{1}{k})$. ■

Proof of Theorem 13: Under the knowledge gradient framework, the incremental amount that one increases one's bid by beyond the expected value of the advertising opportunity is $\delta(E_{\overline{x}'}[U_{k+1}(\overline{x}')] - U_k(\overline{x}))$. And under the full dynamic programming problem, the incremental amount that one increases one's bid is $\delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$. Thus if ΔV_{k+1} denotes the difference between the values of $E_{\overline{x}'}[V_{k+1}(\overline{x}')]$ and $E_{\overline{x}'}[U_{k+1}(\overline{x}')]$ and ΔV_k denotes the difference between the values of $V_k(\overline{x})$ and $U_k(\overline{x})$, then the difference between the incremental amount that one increases one's bid under the full dynamic programming problem and under the knowledge gradient framework is $\delta(\Delta V_{k+1} - \Delta V_k)$.

But a condition of the theorem is that $\Delta V_{k+1} < \Delta V_k$. Thus $\delta(\Delta V_{k+1} - \Delta V_k) < 0$, and the difference between the incremental amount that one increases one's bid under the full dynamic programming problem and the incremental amount that one increases one's bid under the knowledge gradient framework is negative. From this it follows that the incremental amount by which one would increase one's bid under the knowledge gradient framework is indeed greater than it is under the full dynamic programming problem. ■

Proof of Theorem 14: In the knowledge gradient framework, $U_k(\overline{x})$ is just the discounted sum of the value of simply bidding $z = \overline{x}$ in each period when an ad has received k impressions so far and one's best estimate for the eCPM of the ad is $\overline{x}$. Now we know from applying Lemma 4 to the special case in which $z = \overline{x}$ that the per-period payoff from bidding $z = \overline{x}$ in each period when an ad has received k impressions so far and one's best estimate for the eCPM of the ad is $\overline{x}$ is $-\frac{1}{2}\sigma_k^2 f(\overline{x}) - a(\overline{x})\sigma_k^4 + o(\sigma_k^4) = -\frac{1}{2k}s^2(\overline{x})f(\overline{x}) + O(\frac{1}{k^2})$. Thus $U_k(\overline{x}) = -\frac{1}{2(1-\delta)k}s^2(\overline{x})f(\overline{x}) + O(\frac{1}{k^2})$.

But we have seen in Theorem 7 that when $V_k(\overline{x}) = -\frac{1}{2(1-\delta)k}s^2(\overline{x})f(\overline{x}) + O(\frac{1}{k^2})$ and the per-period payoff from making the optimal bid is $-\frac{1}{2}\sigma_k^2 f(\overline{x}) - a(\overline{x})\sigma_k^4 + o(\sigma_k^4)$, then it must be the case that $E_{\overline{x}'}[V_{k+1}(\overline{x}'))] - V_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k, where $v(\overline{x}) \equiv \frac{1}{2(1-\delta)}s^2(\overline{x})f(\overline{x})$. An identical argument illustrates that when $U_k(\overline{x}) = -\frac{1}{2(1-\delta)k}s^2(\overline{x})f(\overline{x}) + O(\frac{1}{k^2})$ and the per-period payoff from making the optimal bid is $-\frac{1}{2}\sigma_k^2 f(\overline{x}) - a(\overline{x})\sigma_k^4 + o(\sigma_k^4)$, then it must be the case that $E_{\overline{x}'}[U_{k+1}(\overline{x}'))] - U_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k, where $v(\overline{x}) \equiv \frac{1}{2(1-\delta)}s^2(\overline{x})f(\overline{x})$. The result then follows. ■

Proof of Theorem 16: Since the optimal bid for advertiser i if advertiser i submits the highest eCPM bid amongst the bidders with unknown eCPMs satisfies $z_i = \overline{x}_i + \delta(E_{\overline{x}'(i)}[V_{\vec{k}'(i)}(\overline{x}'(i))] - V_{\vec{k}}(\overline{x}))$, it follows that $E_{x_i}[-\int_{x_i}^{z_i}F(z_i)\,dy + \delta F(z_i)(E_{\overline{x}'(i)}[V_{\vec{k}'(i)}(\overline{x}'(i))] - V_{\vec{k}}(\overline{x}))] = 0$ when advertiser i submits the optimal bid z_i. By substituting this into the equation for the value of the dynamic programming problem $V_{\vec{k}}(\overline{x})$, it follows that the value of this dynamic programming problem is always equal to $\frac{1}{1-\delta}\int_{x_j}^{z_i}F(y)\,dy$, which is an increasing function of z_i. From this it follows that if the optimal bid for advertiser i if advertiser i submits the highest bid of the advertisers with unknown eCPMs is higher than the optimal bid for all other advertisers with unknown eCPMs if one of these other advertisers submits the highest bid of the advertisers with unknown eCPMs, then the decision maker's

payoff from the game is maximized by having advertiser i submit the highest bid of all the advertisers with unknown eCPMs. ■

Proof of Theorem 17: Recall from Lemma 4 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 $-E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = -\int_{\overline{x}}^{z} F(z) - F(y)\ dy - \frac{1}{2}\sigma_k^2 f(\overline{x}) + o(\sigma_k^2)$ for large k. Now if $z = \overline{x} + \frac{c(\overline{x})}{k^\alpha}$ for some constant $c(\overline{x}) \neq 0$, then $\int_{\overline{x}}^{z} F(z) - F(y)\ dy = \int_{\overline{x}}^{\overline{x}+\frac{c(\overline{x})}{k^\alpha}} f(\overline{x})(\overline{x} + \frac{c(\overline{x})}{k^\alpha} - y)\ dy + o(\frac{1}{k^{2\alpha}}) = f(\overline{x})\frac{c^2(\overline{x})}{2k^{2\alpha}} + o(\frac{1}{k^{2\alpha}})$. Thus the auctioneer's per-period payoff if the auctioneer uses a bid for the ad with unknown eCPM of the form $z = \overline{x} + \frac{c(\overline{x})}{k^\alpha}$ is $-\frac{c^2(\overline{x})}{2k^{2\alpha}} f(\overline{x}) - \frac{1}{2}\sigma_k^2 f(\overline{x}) + o(\frac{1}{k^{2\alpha}})$ if $c(\overline{x}) \neq 0$ and $-\frac{1}{2}\sigma_k^2 f(\overline{x}) + o(\sigma_k^2)$ if $c(\overline{x}) = 0$.

Thus if $c(\overline{x}) = 0$, the auctioneer's per-period payoff is $-\frac{1}{2k}s^2(\overline{x})f(\overline{x}) + o(\frac{1}{k})$. We then know from similar reasoning to that in the proof of Theorem 10 that if this is the auctioneer's per-period payoff, then the auctioneer's total payoff from the game is $-\frac{1}{2(1-\delta)k}s^2(\overline{x})f(\overline{x}) + o(\frac{1}{k})$ regardless of the learning rate. Similarly, if $c(\overline{x}) \neq 0$ and $\alpha = \frac{1}{2}$, then the auctioneer's per-period payoff is $-\frac{1}{2k}f(\overline{x})(s^2(\overline{x}) + c^2(\overline{x})) + o(\frac{1}{k})$, and we know from identical reasoning that the auctioneer's total payoff is $-\frac{1}{2(1-\delta)k}f(\overline{x})(s^2(\overline{x}) + c^2(\overline{x})) + o(\frac{1}{k})$, which is strictly less than the auctioneer's total payoff from the game when $c(\overline{x}) = 0$ for sufficiently large k.

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

Observation 23 *Suppose the auctioneer displays the ad with the highest eCPM bid with probability $1 - \epsilon$ and displays an ad uniformly at random with probability $\epsilon > 0$. Then the optimal constant ϵ for such an algorithm is $\epsilon = 0$ for sufficiently large k.*

Proof Recall from Lemma 4 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 $-E_\epsilon[\int_{\overline{x}+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = -\int_{\overline{x}}^{z} F(z) - F(y)\ dy - \frac{1}{2}\sigma_k^2 f(\overline{x}) + o(\sigma_k^2)$ for large k. Note that displaying an ad uniformly at random is equivalent to making a bid of 0 for the ad with unknown eCPM with probability $\frac{1}{2}$ and making a bid of ∞ for the ad with unknown eCPM with probability $\frac{1}{2}$. Since $\int_{\overline{x}}^{z} F(z) - F(y)\ dy > 0$ for either $z = 0$ or $z = \infty$, it follows that the auctioneer's expected per-period payoff if the auctioneer follows the strategy in the statement of the observation is no greater than $-c\epsilon$ for some constant $c > 0$ for large k.

However, if the auctioneer always uses a bid of $z = \overline{x}$ (as would be the case when $\epsilon = 0$), then we know from the proof of Theorem 17 that the auctioneer's per-period payoff is $-\frac{1}{2k}s^2(\overline{x})f(\overline{x}) + o(\frac{1}{k})$ for large k. Thus for sufficiently large k, the auctioneer always

achieves a larger per-period payoff by setting $\epsilon = 0$ than by using any positive value of ϵ, so the optimal constant for this algorithm is $\epsilon = 0$. ■

Proof of Theorem 18: We know from Theorem 7 that $E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}) = \frac{v(\overline{x})}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k, where $v(\overline{x}) = \frac{1}{2(1-\delta)} s^2(\overline{x}) f(\overline{x})$, and we also know from the proof of Theorem 3 that the derivative of the seller's expected payoff from making a bid of z with respect to z is $f(z)(\overline{x} - z + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$. Thus if $\Delta V \equiv E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x})$, then the difference between the auctioneer's expected payoff from making a bid of $\overline{x}$ and a bid of $\overline{x} + \frac{\delta}{2(1-\delta)k(k+1)} s^2(\overline{x}) f(\overline{x})$ is

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

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

Proof of Theorem 19: The theoretically optimal strategy for the auctioneer would entail submitting a bid of $z = \overline{x} + \delta(E_{\overline{x}'}[V_{k+1}(\overline{x}')] - V_k(\overline{x}))$ in each time period. By the same reasoning as in the proof of Theorem 18, we know the difference between the auctioneer's expected payoff from making a bid of $\overline{x}$ and making a bid of $z = \overline{x} + \delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x})] - V_k(\overline{x}))$ is $\frac{f(\overline{x})\delta^2(\Delta V)^2}{2(1-\delta)} + o((\Delta V)^2)$. Since the auctioneer's payoff from using the approximately optimal bidding strategy is also $\frac{f(\overline{x})\delta^2(\Delta V)^2}{2(1-\delta)} + o((\Delta V)^2)$, it follows that the difference between the auctioneer's payoff under the approximately optimal bidding strategy and the maximum possible payoff the auctioneer could obtain theoretically is $o((\Delta V)^2) = o(\frac{1}{k^4})$.

But we know from Theorem 18 that the difference between the auctioneer's payoff under the approximately optimal bidding strategy and the greedy strategy is $\frac{\delta^2}{8(1-\delta)^3 k^4} s^4(\overline{x}) f^2(\overline{x}) + o(\frac{1}{k^4})$. Thus 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 of Theorem 20: A consequence of Theorem 13 is that the difference between the auctioneer's payoff under the theoretically optimal strategy and the auctioneer's payoff from the greedy strategy is no greater than the difference between the auctioneer's payoff from the theoretically optimal strategy and the auctioneer's payoff under the greedy strategy when no learning is possible in future periods. Thus we seek to bound the difference between the auctioneer's payoff from the theoretically optimal strategy and the auctioneer's payoff under the greedy strategy when no learning is possible in future periods.

Let α and β denote the parameters of the beta distribution. If no learning ever took place and the auctioneer followed the greedy strategy, then in all periods the auctioneer would show the highest competing bidder if this bidder had an eCPM bid p satisfying $p > \frac{\alpha}{\alpha+\beta}$ and show the bidder with unknown eCPM otherwise. If the auctioneer showed the ad with unknown eCPM in the first period, this ad received a click, and no learning

took place in future periods, then in future periods the auctioneer would show the highest competing bidder if and only if this bidder had an eCPM bid p satisfying $p > \frac{\alpha+1}{\alpha+\beta+1}$. And if the auctioneer showed the ad with unknown eCPM, this ad did not receive a click, and no learning took place in future periods, then in future periods the auctioneer would show the highest competing bidder if and only if this bidder had an eCPM bid p satisfying $p > \frac{\alpha}{\alpha+\beta+1}$.

From this it follows that if no learning takes place in future periods, then the auctioneer's payoff in any given period in the future is guaranteed to be the same regardless of whether the ad with unknown eCPM was shown in the first period if either $p > \frac{\alpha+1}{\alpha+\beta+1}$ or $p < \frac{\alpha}{\alpha+\beta+1}$ in that particular period. The only circumstances under which the auctioneer's expected payoff in a future period t will differ as a result of showing the ad with unknown eCPM in the first period is if this ad receives a click in the first period and $p \in (\frac{\alpha}{\alpha+\beta}, \frac{\alpha+1}{\alpha+\beta+1})$ in period t or if the ad does not receive a click in the first period and $p \in (\frac{\alpha}{\alpha+\beta+1}, \frac{\alpha}{\alpha+\beta})$ in period t. In the first case, the auctioneer's payoff in period t as a result of showing the ad with unknown eCPM in the first period exceeds the auctioneer's payoff under normal circumstances by $\frac{\alpha+1}{\alpha+\beta+1} - p$, and in the second case the auctioneer's payoff in period t as a result of showing the ad with unknown eCPM in the first period exceeds the auctioneer's payoff under normal circumstances by an amount $p - \frac{\alpha}{\alpha+\beta+1}$.

Now the probability the ad with unknown eCPM receives a click in the first period if this ad is shown is $\frac{\alpha}{\alpha+\beta}$ and the probability this ad does not receive a click in the first period if this ad is shown is $\frac{\beta}{\alpha+\beta}$. By combining this with the result in the previous paragraph, it follows that the maximum possible expected payoff difference that the auctioneer can obtain from future periods as a result of showing the ad with unknown eCPM in the first period is $\frac{\delta}{1-\delta}[\frac{\alpha}{\alpha+\beta}\int_{\frac{\alpha}{\alpha+\beta}}^{\frac{\alpha+1}{\alpha+\beta+1}}(\frac{\alpha+1}{\alpha+\beta+1} - p)\overline{f}\, dp + \frac{\beta}{\alpha+\beta}\int_{\frac{\alpha}{\alpha+\beta+1}}^{\frac{\alpha}{\alpha+\beta}}(p - \frac{\alpha}{\alpha+\beta+1})\overline{f}\, dp]$, where $\overline{f} \equiv \sup_p f(p)$. This payoff difference equals $\frac{\delta\overline{f}}{2(1-\delta)}[\frac{\alpha}{\alpha+\beta}(\frac{\alpha+1}{\alpha+\beta+1} - \frac{\alpha}{\alpha+\beta})^2 + \frac{\beta}{\alpha+\beta}(\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta+1})^2] = \frac{\delta\overline{f}}{2(1-\delta)}[\frac{\alpha}{\alpha+\beta}(\frac{\beta}{(\alpha+\beta)(\alpha+\beta+1)})^2 + \frac{\beta}{\alpha+\beta}(\frac{\alpha}{(\alpha+\beta)(\alpha+\beta+1)})^2] = \frac{\delta\overline{f}}{2(1-\delta)}\frac{\alpha\beta(\alpha+\beta)}{(\alpha+\beta)^3(\alpha+\beta+1)^2} = \frac{\delta\overline{f}}{2(1-\delta)}\frac{(\alpha+\beta)^2\alpha^2\beta^2}{\alpha\beta(\alpha+\beta)^4(\alpha+\beta+1)^2}$.

Now the expected value for a beta distribution is $\frac{\alpha}{\alpha+\beta}$ and the variance in a beta distribution is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Thus since the bidder's expected eCPM is ω and the standard deviation in the bidder's expected eCPM is $\gamma\omega$, it follows that $\frac{\alpha+\beta}{\alpha} = \frac{1}{\omega}$, $\frac{\alpha+\beta}{\beta} = \frac{1}{1-\omega}$, and $\frac{\alpha^2\beta^2}{(\alpha+\beta)^4(\alpha+\beta+1)^2} = \gamma^4\omega^4$. From this it follows that $\frac{\delta\overline{f}}{2(1-\delta)}\frac{(\alpha+\beta)^2\alpha^2\beta^2}{\alpha\beta(\alpha+\beta)^4(\alpha+\beta+1)^2} = \frac{\delta\overline{f}}{2(1-\delta)}\frac{\gamma^4\omega^3}{1-\omega}$. Thus the maximum additional payoff increase that one can obtain from future periods as a result of showing the ad with uncertain eCPM in the first period is no greater than $\frac{\delta\overline{f}}{2(1-\delta)}\frac{\gamma^4\omega^3}{1-\omega}$.

Now if ΔV denotes the change in payoff that one obtains from future periods as a result of showing the ad with uncertain eCPM in the first period, then the value of showing the ad with uncertain eCPM in the first period is $\omega + \Delta V$. Thus the theoretically optimal strategy will specify a bid of $\omega + \Delta V$ for the bidder with uncertain eCPM, whereas the greedy strategy will specify a bid of ω, so the theoretically optimal strategy will only show a different ad when the highest competing eCPM bid, p, satisfies $p \in [\omega, \omega + \Delta V]$. Furthermore, in the cases where the theoretically optimal strategy specifies a different bid, the theoretically optimal strategy achieves a payoff that exceeds that of the greedy strategy by an amount $\omega + \Delta V - p$, where p denotes the highest competing eCPM bid. From this it follows that the

difference in expected payoff that one obtains as a result of using the theoretically optimal strategy rather than the greedy strategy is no greater than $\frac{1}{1-\delta}\int_{\omega}^{\omega+\Delta V}(\omega+\Delta V-p)f(p)\,dp$.

Thus if $\overline{f} \equiv \sup_p f(p)$, this payoff difference is no greater than $\frac{\overline{f}}{1-\delta}\int_{\omega}^{\omega+\Delta V}(\omega+\Delta V-p)\,dp = \frac{\overline{f}}{1-\delta}\frac{(\Delta V)^2}{2}$. Thus the difference between the auctioneer's payoff under the theoretically optimal strategy and the auctioneer's payoff from the greedy strategy is no greater than $\frac{\overline{f}}{1-\delta}\frac{(\Delta V)^2}{2}$.

But we have seen earlier that the maximum additional payoff increase that one can obtain from future periods as a result of showing the ad with uncertain eCPM in the first period is no greater than $\frac{\delta\overline{f}}{2(1-\delta)}\frac{\gamma^4\omega^3}{1-\omega}$. Thus we know that $\Delta V \leq \frac{\delta\overline{f}}{2(1-\delta)}\frac{\gamma^4\omega^3}{1-\omega}$. By combining this with the result in the previous paragraph, we see that 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 $\frac{\delta^2\gamma^8\omega^6\overline{f}^3}{8(1-\delta)^3(1-\omega)^2}$. ■

References

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

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

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

N. Anthonisen. On learning to cooperate. *Journal of Economic Theory*, 107(2):253-287, 1993.

N. Anthonisen. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785-2836, 2010.

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.

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

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

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

J.S. Banks and R.K. Sundaram. Denumerable-armed bandits. *Econometrica*, 60(5):1071-1096, 2004.

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.

D. Bergemann and J. Välimäki. Experimentation in markets. *Review of Economic Studies* 67(2):213-234, 2000.

D. Bergemann and J. Välimäki. Learning and strategic pricing. *Econometrica*, 64(5):1125-1149, 1996.

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

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

P. Bolton and C. Harris. Strategic experimentation. *Econometrica* 67(2):349-374, 1999.

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

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

S. Callander and P. Hummel. Preemptive policy experimentation. *Econometrica* 82(4):1509-1528, 2014.

S.E. Chick and N. Gans. Economic analysis of simulation selection problems. *Management Science* 55(3):421-437, 2009.

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

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

P. Frazier, W. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* 21(4):599-613, 2009.

D. Gale. What have we learned from social learning? *European Economic Review* 40(3-5):617-628, 2011.

D. Gale and R.W. Rosenthal. Experimentation, imitation, and stochastic stability. *Journal of Economic Theory* 84(1):1-40, 2011.

A. Ghate. Optimal minimum bids and inventory scrapping in sequential, single-unit, Vickrey auctions with demand learning. *European Journal of Operations Research* 245(2):555-570, 2015.

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

E. Hazan and S. Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research* 12:1287-1311, 2011.

K. Iyer, R. Johari, and M. Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science* 60(12):2949-2970, 2014.

S.M. Kakade, I. Lobel, and H. Nazerzadeh. Optimal dynamic mechanism design and the virtual pivot mechanism. *Operations Research* 61(4):837-854, 2013.

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

G. Keller and S. Rady. Strategic experimentation with Poisson bandits. *Theoretical Economics* 5(2):275-311, 2010.

G. Keller, S. Rady, and M. Cripps. Strategic experimentation with exponential bandits. *Econometrica* 73(1):39-68, 2010.

S. Lahaie and R.P. McAfee. Efficient ranking in sponsored search. In *Proceedings of the* 7^{th} *International Workshop on Internet and Network Economics* (WINE), pages 254-265. Springer, 2011.

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

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

S. Mannor and J.N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5:623-648, 2004.

B.C. May, N. Korda, A. Lee, and D.S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 13(1): 2069-2106, 2012.

R.P. McAfee. The design of advertising exchanges. *Review of Industrial Organization* 39(3):169-185, 2011.

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

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

M. Ostrovsky and M. Schwarz. Reserve prices in Internet advertising auctions: a field experiment. Stanford University Typescript, 2009.

A. Pavan, I. Segal, and J. Toikka. Dynamic mechanism design: a Myersonian approach. *Econometrica* 82(2):601-653, 2014.

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

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.

I.O. Ryzhov, P.I. Frazier, and W.B. Powell. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. *Procedia Computer Science* 1:1629-1639, 2010.

I.O. Ryzhov, W.B. Powell, and P.I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research* 60(1):180-195, 2012.

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.

A. Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research* 15(1):2533-2568, 2014.

B. Strulovici. Learning while voting: determinant of collective experimentation. *Econometrica* 78(3):933-971, 2010.

H.R. Varian. Online ad auctions. *American Economic Review: Papers & Proceedings* 99(2):430-434, 2009.

X. Vives. Learning from others: a welfare analysis. *Games and Economic Behavior* 20(2):177-200, 1997.

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

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