Dynamic Cost-Per-Action Mechanisms and Applications to Online Advertising

Hamid Nazerzadeh
Stanford University
Stanford, CA 94304
hamidnz@stanford.edu

Amin Saberi
Stanford University
Stanford, CA 94304
saberi@stanford.edu

Rakesh Vohra
Northwestern University
Evanston, IL 60208
r-vohra@kellogg.nwu.edu

ABSTRACT

We study the Cost-Per-Action or Cost-Per-Acquisition (CPA) charging scheme in online advertising. In this scheme, instead of paying per click, the advertisers pay only when a user takes a specific action (e.g. fills out a form) or completes a transaction on their websites.

We focus on designing efficient and incentive compatible mechanisms that use this charging scheme. We describe a mechanism based on a sampling-based learning algorithm that under suitable assumptions is asymptotically individually rational, asymptotically Bayesian incentive compatible and asymptotically ex-ante efficient.

In particular, we demonstrate our mechanism for the case where the utility functions of the advertisers are independent and identically-distributed random variables as well as the case where they evolve like independent reflected Brownian motions.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Economics, Algorithm, Theory

Keywords

Mechanism Design, Cost-Per-Action, Internet Advertising

1. INTRODUCTION

Currently, the main two charging models in the online advertising industry are cost-per-impression (CPM) and cost-per-click (CPC). In the CPM model, the advertisers pay the publisher for the impression of their ads. CPM is commonly used in traditional media (e.g. magazines and television) or banner advertising and is more suitable when the goal of the advertiser is to increase brand awareness.

A more attractive and more popular charging model in online advertising is the CPC model in which the advertisers pay the publisher only when a user clicks on their ads. In the last few years, there has been a tremendous shift towards the CPC charging model. CPC is adopted by search engines such as Google or Yahoo! for the placement of ads next to search results (also known as sponsored search) and on the website of third-party publishers.

In this paper we will focus on another natural and widely advocated charging scheme known as the Cost-Per-Action or Cost-Per-Acquisition (CPA) model. In this model, instead of paying per click, the advertiser pays only when a user takes a specific action (e.g. fills out a form) or completes a transaction. Recently, several companies like Google, eBay, Amazon, Advertising.com, and Snap.com have started to sell advertising in this way.

CPA models can be the ideal charging scheme, especially for small and risk averse advertisers. We will briefly describe a few advantages of this charging scheme over CPC and refer the reader to [18] for a more detailed discussion.

One of the drawbacks of the CPC scheme is that it requires the advertisers to submit their bids before observing the profits generated by the users clicking on their ads. Learning the expected value of each click, and therefore the right bid for the ad, is a prohibitively difficult task especially in the context of sponsored search in which the advertisers typically bid for thousands of keywords. CPA eliminates this problem because it allows the advertisers to report their payoff *after* observing the user's action.

Another drawback of the CPC scheme is its vulnerability to click fraud. Click fraud refers to clicks generated by someone or something with no genuine interest in the advertisement. Such clicks can be generated by the publisher of the content who has an interest in receiving a share of the revenue of the ad or by a rival who wishes to increase the cost of advertising for the advertiser. Click fraud is considered by many experts to be the biggest challenge facing the online advertising industry [13, 10, 23, 20]. CPA schemes are less vulnerable because generating a fraudulent action is typically more costly than generating a fraudulent click. For example, an advertiser can define the action as a sale and pay the publisher only when the ad yields profit[1].

On the other hand, there is a fundamental difference between CPA and CPC charging models. A click on the ad can be observed by both advertiser and publisher. However, the action of the user is hidden from the publisher and is observable only by the advertiser. Although the publisher can require the advertisers to install a software that will monitor actions that take place on their web site, even moderately sophisticated advertisers can find a way to manipulate the software if they find it sufficiently profitable.

[1] CPA makes generating a fraudulent action a more costly enterprize, but not impossible (e.g., using a stolen credit).

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Are the publishers exposed to the manipulation or misreporting of the advertisers in the CPA scheme? Does CPA create an incentive for the advertisers to misreport the number of actions or their payoffs for the actions? The main result of this paper is to give a negative answer to these questions. We design a mechanism that, asymptotically and under reasonable assumptions, removes the incentives of the advertisers to misreport their payoffs. At the same time, our mechanism has the same asymptotic efficiency and hence revenue as the currently used CPC mechanisms. We will use techniques in learning and mechanism design to obtain this result.

In the next section, we will formally describe our model in mechanism design terminology (see [21].) We will refer to advertisers as agents and to the impression of an ad as an item. For simplicity of exposition only, we assume only one advertisement slot per page. In section 6 we outline how to extend our results to the case where more than one advertisement can be displayed in each page. Although our work is essentially motivated by online advertising, we believe that the application of our mechanism is not limited this domain.

1.1 Model

We study the following problem: there are a number of self-interested agents competing for identical items sold repeatedly at times $t = 1, 2, \cdots$. At each time t, a mechanism allocates the item to one of the agents. Agents *discover* their utility for the good only if it is allocated to them. If agent i receives the good at time t, she discovers utility u_{it} (denominated in money) for it and reports (not necessarily truthfully) the realized utility to the mechanism. Then, the mechanism determines how much the agent has to pay for receiving the item. We allow the utility of an agent to change over time. For this environment we are interested in auction mechanisms which have the following four properties.

1. The mechanism is individually rational in each period.
2. Agents have an incentive to truthfully report their realized utilities.
3. The efficiency (and revenue) is, in an appropriate sense, not too small compared to a second price auction.
4. The correctness of the mechanism does not depend on an a-priori knowledge of the distribution of u_{it}'s. This feature is motivated by the Wilson doctrine [24] [2].

The precise manner in which these properties are formalized is described in section 2.

We will build our mechanisms on a sampling-based learning algorithm. The learning algorithm is used to estimate the expected utility of the agents, and consists of two *alternating* phases: exploration and exploitation. During an exploration phase, the item is allocated for free to a randomly chosen agent. During an exploitation phase, the mechanism allocates the item to the agent with the highest estimated expected utility. After each allocation, the agent who has received the item, discovers her utility and reports it to the mechanism. Subsequently, the mechanism updates the estimate of utilities and determines the payment.

We characterize a class of learning algorithms that ensure that the corresponding mechanism has the four desired properties. The main difficulty in obtaining this result is the following: since there is uncertainty about the utilities, it is possible that in some periods the item is allocated to an agent who does not have the highest utility in that period. Hence, the natural second-highest price payment rule would violate individual rationality. On the other hand, if the mechanism does not charge an agent because her reported utility after the allocation is low, it gives her an incentive to shade her reported utility down. Our mechanism solves these problems by using an adaptive, cumulative pricing scheme.

We illustrate our results by identifying simple mechanisms that have the desired properties. We demonstrate these mechanisms for the case in which the u_{it}'s are independent and identically-distributed random variables as well as the case where their expected values evolve like independent reflected Brownian motions. In these cases the mechanism is actually *ex-post* individually rational.

In our proposed mechanism, the agents do not have to bid for the items. This is advantageous when the bidders themselves are unaware of their utility values. However, in some cases, an agent might have a better estimate of her utility for the item than our mechanism. For this reason, we describe how we can slightly modify our mechanism to allow those agents to bid directly.

1.2 Related Work

There is a large number of interesting results on using machine learning techniques in mechanism design. We only briefly survey the main techniques and ideas and compare them with the approach of this paper.

Most of these works, like [5, 8, **?**, 17], consider one-shot games or repeated auctions in which the agents leave the environment after they received an item. In our setting we may allocates items to an agent several times and hence, we need to consider the strategic behavior of the agents over time. There is also a big literature on regret minimization or expert algorithms. In our context, these algorithms are applicable even if the utilities of the agents are changing arbitrarily. However, the efficiency (and therefore the revenue) of these algorithms is comparable to the mechanisms that allocates the item to the single best agent (expert) (e.g. see [16]). Our goal is more ambitious: our efficiency is close the most efficient allocation which might allocate the item to different agents at different times. On the other hand, we focus on utility values that change smoothly (e.g. like a Brownian motion).

In a finitely repeated version of the environment considered here, Athey and Segal [2] construct an efficient, budget balanced, mechanism where truthful revelation in each period is Bayesian incentive compatible. Bapna and Weber [4] consider the infinite horizon version of [2] and propose a class of incentive compatible mechanisms based on the Gittins index (see [11]). Taking a different approach, Bergemann and Välimäki [6] and Cavallo et al. [9] propose an incentive compatible generalization of the Vickrey-Clark-Groves mechanism based on the marginal contribution of each agent for this environment. All these mechanisms need the exact solution of the underlying optimization problems, and therefore require complete information about the prior of the utilities

[2] Wilson criticizes relying too much on common-knowledge assumptions.

of the agents; also, they do not apply when the evolution of the utilities of the agents is not stationary over time. This violates the last of our desiderata. For a comprehensive survey in dynamic mechanism design literature see [22].

In the context of sponsored search, attention has focused on ways of estimating click through rates. Gonen and Pavlov [12] give a mechanism which learns the click-through rates via sampling and show that truthful bidding is, with high probability, a (weakly) dominant strategy in this mechanism. Along this line, Wortman et al. [25] introduced an exploration scheme for learning advertisers' click-through rates in sponsored search which maintains the equilibrium of the system. In these works, unlike ours, the distribution of the utilities of agents are assumed to be fixed over time.

Immorlica et al. [14], and later Mahdian and Tomak [18], examine the vulnerability of various procedures for estimating click through, and identify a class of click through learning algorithms in which fraudulent clicks cannot increase the expected payment per impression by more than $o(1)$. This is under the assumption that the slot of an agent is fixed and the bids of other agents remain constant overtime. In contrast, we study conditions which guarantee incentive compatibility and efficiency, while the utility of (all) agents may evolve over time.

2. DEFINITIONS AND NOTATION

Suppose n agents competing in each period for a single item. The item is sold repeatedly at time $t = 1, 2, \cdots$. Denote by u_{it} the nonnegative utility of agent i for the item at time t. Utilities are denominated in a common monetary scale.

The utilities of agents may evolve over time according to a stochastic process. We assume that for $i \neq j$, the evolution of u_{it} and u_{jt} are independent stochastic processes. We also define $\mu_{it} = E[u_{it}|u_{i1}, \cdots, u_{i,t-1}]$. Throughout this paper, expectations are taken conditioned on the complete history. For simplicity of notation, we now omit those terms that denote such a conditioning. With notational convention, it follows, for example, that $E[u_{it}] = E[\mu_{it}]$. Here the second expectation is taken over all possible histories.

Let $\mathcal{M}$ be a mechanism used to sell the items. At each time, $\mathcal{M}$ allocates the item to one of the agents. Let i be the agent who has received the item at time t. Define x_{it} to be the variable indicating the allocation of the item to i at time t. After the allocation, agent i observes her utility, u_{it}, and then reports r_{it}, as her utility for the item, to the mechanism. Note that we do not require an agent to know her utility for possessing the item in advance of acquiring it. The mechanism then determines the payment, denoted by p_{it}.

Definition 1. *An agent i is* truthful *if $r_{it} = u_{it}$, for all time $x_{it} = 1, t > 0$.*

Our goal is to design a mechanism which has the following properties. We assume n, the number of agents, is constant.

Individual Rationality: A mechanism is *ex-post* individually rational if for any time $T > 0$ and any agent $1 \leq i \leq n$, the total payment of agent i does not exceed the sum of her reports:

$$\sum_{t=1}^{T} x_{it} r_{it} - p_{it} > 0.$$

M is *asymptotically ex-ante individually rational* if:

$$\liminf_{T\to\infty} E[\sum_{t=1}^{T} x_{it}\mu_{it} - p_{it}] \geq 0.$$

Incentive Compatibility: This property implies that truthfulness defines an asymptotic Bayesian Nash equilibrium. Consider agent i and suppose all agents except i are truthful. Let $U_i(T)$ be the expected total profit of agent i, if agent i is truthful between time 1 and T. Also, let $\widetilde{U}_i(T)$ be the maximum of expected profit of agent i under any other strategy. *Asymptotic incentive compatibility* requires that

$$\widetilde{U}_i(T) - U_i(T) = o(U_i(T)).$$

Efficiency: An ex-ante efficient mechanism allocates the item to an agent in $\operatorname{argmax}_i\{\mu_{it}\}$ at each time t (and for each history). The total social welfare obtained by an ex-ante efficient mechanism up to time T is $E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}]$. Let $W(T)$ be the expected welfare of mechanism $\mathcal{M}$ between time 1 and T, when all agents are truthful, i.e.,

$$W(T) = E[\sum_{t=1}^{T}\sum_{i=1}^{n} x_{it}\mu_{it}]$$

Then, $\mathcal{M}$ is *asymptotically ex-ante efficient* if:

$$E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - W(T) = o(W(T)).$$

3. PROPOSED MECHANISM

We build our mechanism on top of a learning algorithm that estimates the expected utility of the agents. We refrain from an explicit description of the learning algorithm. Rather, we describe sufficient conditions for a learning algorithm that can be extended to a mechanism with all the properties we seek (see section 3.1). In section 4 and 5 we give two examples of environments where learning algorithms satisfying these sufficient conditions exist.

The mechanism consists of two phases: *explore* and *exploit*. During the explore phase, with probability $\eta(t)$, $\eta : \mathbb{N} \to [0, 1]$, the item is allocated for free to a randomly chosen agent. During the exploit phase, the mechanism allocates the item to the agent with the highest estimated expected utility. Afterwards, the agent reports her utility to the mechanism and the mechanism determines the payment. We first formalize our assumptions about the learning algorithm and then we discuss the payment scheme. The mechanism is given in Figure 1.

The learning algorithm, samples u_{it}'s at rate $\eta(t)$, and based on the history of the reports of agent i, returns an estimate of μ_{it}. Let $\widehat{\mu}_{it}(T)$ be the estimate of the algorithm for μ_{it} conditional on the history of the reports up to time T. The history of the reports of agent i up to time T is the sequence of the reported values and times of observation of u_{it} up to but not including time T. Note that we allow $T > t$. Thus, information at time $T > t$ can be used to revise an estimate of μ_{it} made at some earlier time. We assume that increasing the number of samples only increases the

```
For $t = 1, 2, \ldots$
    With probability $\eta(t)$, *explore*:
        Uniformly at random, allocate the item to an agent $i$, $1 \leq i \leq n$.
        $p_{it} \leftarrow 0$
    With probability $1 - \eta(t)$, *exploit*:
        Randomly allocate the item to an agent $i \in \text{argmax}_i\{\widehat{\mu}_{it}(t)\}$.
        $p_{it} \leftarrow \sum_{k=1}^{t-1} y_{ik} \min\{\widehat{\gamma}_k(t), \widehat{\mu}_{ik}(k)\} - \sum_{k=1}^{t-1} p_{ik}$
    $r_{it} \leftarrow$ the report of agent $i$.
    $p_{jt} \leftarrow 0$, $j \neq i$
```

Figure 1: Mechanism $\mathcal{M}$

accuracy of the estimations, i.e. for any truthful agent i, and times $T_1 \leq T_2$:

$$E[|\widehat{\mu}_{it}(T_1) - \mu_{it}|] \quad \geq \quad E[|\widehat{\mu}_{it}(T_2) - \mu_{it}|]. \qquad (1)$$

In the inequality above, and in the rest of the paper, the expectations of $\widehat{\mu}_{it}$ are taken over the evolution of u_{it}'s and the random choices of the mechanism. For simplicity of notation, we omit those terms that denote such a conditioning.

To describe the payments recall that γ_t is the second highest μ_{it} and let $\widehat{\gamma}_t(T) = \max_{j \neq i}\{\widehat{\mu}_{jt}(T)\}$, where i is the agent who received the item at time t. We define y_{it} to be the indicator variable of the allocation of the item to agent i during an exploit phase. The payment of agent i at time t, denoted p_{it}, is determined so that:

$$\sum_{k=1}^{t} p_{ik} = \sum_{k=1}^{t-1} y_{ik} \min\{\widehat{\gamma}_k(t), \widehat{\mu}_{ik}(k)\}.$$

An agent only pays for items that are allocated to her during the exploit phase, up to but not including time t. At time t, the payment of agent i for the item she received at time $k < t$ is $\min\{\widehat{\gamma}_k(t), \widehat{\mu}_{ik}(k)\}$. The first term is the reminiscence of the second highest pricing scheme. The second term, under some reasonable conditions, leads to individually rationality. Since the estimations of learning algorithm for the utilities of agents become more precise over time, our adaptive cumulative payment scheme allows it to correct for errors in the past.

3.1 Sufficient Conditions

We start with a condition that guarantees asymptotic ex-ante individual rationality and asymptotic incentive compatibility. Let l_{it} be the last time up to time t that the item is allocated to agent i within an exploit phase. If i has not been allocated any item yet, l_{it} is defined to be zero. Also, define $\Delta_t = \max_i\{|\widehat{\mu}_{it}(t) - \mu_{it}|\}$, assuming all agents were truthful up to time t.

Theorem 1. *If for the learning algorithm, for all* $1 \leq i \leq n$*, and* $T > 0$*:*

$(C1)$ $\quad E[\max_{1 \leq t \leq T}\{\mu_{it}\} + \sum_{t=1}^{T} \Delta_t] = o(E[\sum_{t=1}^{T} \eta(t)\mu_{it}])$

then mechanism $\mathcal{M}$ is asymptotically ex-ante individually rational and incentive compatible.

We outline the proof first. As we prove in Lemma 2, by condition $(C1)$, the expected profit of a truthful agent up to time T is at least $(\frac{1}{n} - o(1))E[\sum_{t=1}^{T} \eta(t)\mu_{it}]$. Also, the expected total error in the estimates of the payments up to time T is bounded by $O(E[\sum_{t=1}^{T} \Delta_t])$. We prove that the total utility an agent could obtain by deviating from the truthful strategy, between time 1 and T, is bounded by $O(\max_{1 \leq t \leq T}\{\mu_{it}\} + E[\sum_{t=1}^{T} \Delta_t])$. Hence, the claim follows by condition $(C1)$.

Similar to other applications of learning algorithms, we can observe a natural trade-off between exploitation and exploration rates in our context: higher exploration rates lead to more accurate estimates of the utilities of the agents, at the cost of efficiency. Condition $(C1)$ provides us with a *lower bound* on the exploration rate.

Lemma 2. *If condition $(C1)$ holds, then the expected profit of a truthful agent i up to time T, $U_i(T)$, is at least:*

$$(\frac{1}{n} - o(1))E[\sum_{t=1}^{T} \eta(t)\mu_{it}].$$

Proof. The items that agent i receives during the explore phase are free. The expected total utility of agent i from these items up to time T is $\frac{1}{n}E[\sum_{t=1}^{T} \eta(t)\mu_{it}]$. Let $C_T = \{t < l_{iT} | y_{it} = 1$, if i is truthful$\}$ be the subset of periods that agent i is charged for the item she received within the period.

$$\begin{aligned} U_i(T) &= E[\sum_{t=1}^{T} x_{it}u_{it} - p_{it}] \\ &= E[\sum_{t \notin C_T} x_{it}u_{it}] + E[\sum_{t \in C_T} u_{it} - p_{it}] \\ &\geq \frac{1}{n}E[\sum_{t=1}^{T} \eta(t)\mu_{it}] \\ &\quad + E[\sum_{t \in C_T} (\mu_{it} - \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\})] \qquad (2) \end{aligned}$$

For $t \in C_T$:

$$\begin{aligned} E[(\mu_{it} - \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\})I(t \in C_T)] & \\ \ge E[(\mu_{it} - \widehat{\mu}_{it}(t))I(t \in C_T)] & \\ \ge -E[|\mu_{it} - \widehat{\mu}_{it}(t)|] & \\ \ge -E[\Delta_t] & \end{aligned}$$

Substituting into inequality (2), by condition $(C1)$:

$$\begin{aligned} U_i(T) & \ge \frac{1}{n}E[\sum_{t=1}^{T}\eta(t)\mu_{it}] - E[\sum_{t=1}^{T}\Delta_t] \\ & = \frac{1}{n}E[\sum_{t=1}^{T}\eta(t)\mu_{it}] - o(E[\sum_{t=1}^{T}\eta(t)\mu_{it}]) \quad (3) \end{aligned}$$

□

Proof of Theorem 1: Lemma 2 yields asymptotic ex-ante individual rationality. We show that truthfulness is asymptotically a best response when all other agents are truthful. Fix an agent i intending to deviate and let $\mathcal{S}$ be the strategy she deviates to. Fixing the evolution of all u_{jt}'s, $1 \le j \le n$, and all random choices of the mechanism, i.e. the steps in the explore phase and the randomly chosen agents, let D_T be the times that i receives the item under strategy $\mathcal{S}$ during the exploit phase before time l_{iT}, i.e. $D_T = \{t < l_{iT} | y_{it} = 1, \text{ if the strategy of } i \text{ is } \mathcal{S}\}$. Similarly, let $C_T = \{t < l_{iT} | y_{it} = 1, \text{ if } i \text{ is truthful}\}$. Also, let $\widehat{\mu}'_{it}$, and $\widehat{\gamma}'_t$ correspond to the estimates of the mechanism when the strategy of i is $\mathcal{S}$. We first bound the expected profit of i, under strategy $\mathcal{S}$, during the exploit phase:

$$\begin{aligned} & E[\sum_{t=1}^{T} y_{it}u_{it} - p_{it}] \\ & \quad \le E[\sum_{t \in D_T} \mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}] + \quad (4) \\ & \quad\quad E[\max_{t \le T}\{\mu_{it}\}] \\ & \quad = E[\sum_{t \in D_T \setminus C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}] + \\ & \quad\quad E[\sum_{t \in D_T \cap C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}] + \\ & \quad\quad E[\max_{t \le T}\{\mu_{it}\}] \quad (5) \end{aligned}$$

The term $E[\max_{t \le T}\{\mu_{it}\}]$ bounds the outstanding payment of agent i; recall that the agent has not paid for the last allocated item.

For time $t \ge 1$, we examine two cases:

1. If $t \in D_T \cap C_T$, then agent i, in expectation, cannot decrease the "current price", $\min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}$, by more than $O(\Delta_t)$:

$$\begin{aligned} \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\} & \ge \min\{\widehat{\gamma}'_t(T), \widehat{\gamma}'_t(t)\} \\ & \ge \gamma_t - \max\{\gamma_t - \widehat{\gamma}'_t(T), \gamma_t - \widehat{\gamma}'_t(t)\} \\ & \ge \gamma_t - (\gamma_t - \widehat{\gamma}'_t(T))^+ - (\gamma_t - \widehat{\gamma}'_t(t))^+ \end{aligned}$$

   where $(z)^+ = \max\{z, 0\}$.

   Recall that $\widehat{\gamma}'_t(T) = \max_{j \neq i}\{\widehat{\mu}'_{it}(T)\}$ and all other agent are truthful. Hence, taking expectation from both sides, by (1):

$$\begin{aligned} & E[\min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}I(t \in D_T \cap C_T)] \\ & \quad \ge E[(\gamma_t - (\gamma_t - \widehat{\gamma}'_t(T))^+ - (\gamma_t - \widehat{\gamma}'_t(t))^+)I(t \in D_T \cap C_T)] \\ & \quad \ge E[\gamma_t I(t \in D_T \cap C_T)] - E[2\Delta_t] \quad (6) \end{aligned}$$

2. If $t \in D_T \setminus C_T$, agent i cannot increase her "expected profit", $\mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\}$, by more than $O(\Delta_t)$:

$$\begin{aligned} \mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\} & \le \mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\gamma}'_t(t)\} \\ & \le (\mu_{it} - \widehat{\mu}_{it}(t)) + (\widehat{\mu}_{it}(t) - \gamma_t) \\ & \quad + \max\{\gamma_t - \widehat{\gamma}'_t(T), \gamma_t - \widehat{\gamma}'_t(t)\} \\ & \le 2\Delta_t + (\gamma_t - \widehat{\gamma}'_t(T))^+ \\ & \quad + (\gamma_t - \widehat{\gamma}'_t(t))^+ \end{aligned}$$

   Taking expectation from both sides, by (1):

$$\begin{aligned} & E[(\mu_{it} - \min\{\widehat{\gamma}'_t(T), \widehat{\mu}'_{it}(t)\})I(t \in D_T - C_T)] \\ & \quad \le E[2\Delta_t I(t \in D_T - C_T)] + \\ & \quad\quad E[((\gamma_t - \widehat{\gamma}'_t(T))^+ + (\gamma_t - \widehat{\gamma}'_t(t))^+)I(t \in D_T - C_T)] \\ & \quad \le E[4\Delta_t] \quad (7) \end{aligned}$$

Substituting inequalities (6) and (7) into (5):

$$\begin{aligned} E[\sum_{t=1}^{T} y_{it}u_{it} - p_{it}] & \le E[\sum_{t=1}^{T} 6\Delta_t] + E[\max_{t \le T}\{\mu_{it}\}] \\ & \quad + E[\sum_{t \in D_T \cap C_T} \mu_{it} - \gamma_t] \\ & \le E[\sum_{t=1}^{T} 6\Delta_t] + E[\max_{t \le T}\{\mu_{it}\}] + \quad (8) \\ & \quad E[\sum_{t \in C_T} \mu_{it} - \gamma_t] - E[\sum_{t \in C_T \setminus D_T} \mu_{it} - \gamma_t] \end{aligned}$$

For $t \in C_T$, since $\widehat{\mu}_{it}(t) \ge \widehat{\gamma}_t(t)$, we have:

$$E[\gamma_t - \mu_{it}] \le E[2\Delta_t]$$

Substituting into (8):

$$\begin{aligned} E[\sum_{t=1}^{T} y_{it}u_{it} - p_{it}] & \le 8E[\sum_{t=1}^{T} \Delta_t] + E[\max_{t \le T}\{\mu_{it}\}] \\ & \quad + E[\sum_{t \in C_T} \mu_{it} - \gamma_t] \end{aligned}$$

With algebraic manipulation, using (1), we get:

$$\begin{aligned} E[\sum_{t=1}^{T} y_{it}u_{it} - p_{it}] & \le O(E[\sum_{t=1}^{T} \Delta_t] + E[\max_{t \le T}\{\mu_{it}\}]) \\ & \quad + E[\sum_{t \in C_T} \mu_{it} - \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\}] \end{aligned}$$

By condition (C1), we get the inequality below which completes the proof:

$$\begin{aligned} E[\sum_{t=1}^{T} y_{it}u_{it} - p_{it}] & \le o(E[\sum_{t=1}^{T} \eta(t)\mu_{it}]) \\ & \quad + E[\sum_{t \in C_T} \mu_{it} - \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\}] \end{aligned}$$

and the last inequality follows by $(C1)$. The expected utility of the truthful strategy and $\mathcal{S}$ during the explore phase is equal. Therefore, by Lemma 2, the mechanism is asymptotically incentive compatible. □

In the next theorem we show if the loss in efficiency during exploration asymptotically goes to zero, then by Condition $(C1)$ the mechanism is asymptotically ex-ante efficient.

THEOREM 3. *If for the learning algorithm, in addition to* $(C1)$*, the following condition holds*

$$(C2) \quad E[\textstyle\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}] = o(E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}])$$

then, $\mathcal{M}$ *is asymptotically ex-ante efficient.*

PROOF. $\mathcal{M}$ may fail to be ex-ante efficient for two reasons. First one is the loss in welfare during the exploration when the item is allocated randomly to one of advertisers. The expected loss in this case is equal to $E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]$. Another reason is the mistakes during exploitation. The error in estimation can lead to allocation to an agent who does not value the item the most. At time t, in the worst case, the item might be allocated to an agent whose expected utility is at most $2\Delta_t$ less than the highest expected utility. Therefore, the expected efficiency loss during exploration is bounded by $O(E[\sum_{t=1}^{T} \Delta_t])$. Since, for the expected welfare of $\mathcal{M}$ between time 1 and T, denoted by $W(T)$, we have:

$$\begin{aligned} &E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - W(T) \\ &\qquad = O(E[\sum_{t=1}^{T} (\Delta_t + \eta(t) \max_i\{\mu_{it}\})]) \qquad (9) \end{aligned}$$

But, condition $(C1)$ implies:

$$\begin{aligned} E[\sum_{t=1}^{T} \Delta_t] &= o(E[\sum_{t=1}^{T}\sum_{i=1}^{n} \eta(t)\mu_{it}\}]) \\ &= \theta(E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]) \end{aligned}$$

Plugging into (9):

$$\begin{aligned} E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - W(T) &= O(E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]) \\ &= o(W(T)) \end{aligned}$$

The last equality is followed by $(C2)$ and implies asymptotic ex-ante efficiency. □

While Condition $(C1)$ gives a lower bound on the exploration rate, Condition $(C2)$ gives an *upper bound*. In the next section, we will show with two examples how conditions $(C1)$ and $(C2)$ can be used to adjust the exploration rate of a learning algorithm in order to obtain efficiency and incentive compatibility.

REMARK 1. *In Theorem 3 we showed that under some assumptions, the welfare obtained by the mechanism is asymptotically equivalent to efficient mechanism that every time allocates the item to the agent with the highest expected utility. We can give similar conditions to* $(C2)$ *to guarantee that the revenue of the mechanism is also asymptotically equal to the revenue of the efficient mechanism that every time charges the winning agent the second highest expected utility. To avoid repetition, we refrain from explaining this condition in details.*

3.2 Allowing agents to bid

In mechanism $\mathcal{M}$ no agent explicitly bids for an item. Whether an agent receives an item or not depends on the history of their reported utilities and the estimates that $\mathcal{M}$ forms from them. This may be advantageous when the bidders themselves are unaware of what their utilities will be. However, when agents may posses a better estimate of their utilities we would like to make use of that. For this reason we describe how to modify $\mathcal{M}$ so as to allow agents to bid for an item.

If time t occurs during an exploit phase let $\mathcal{B}_t$ be the set of the agents who bid at this time. The mechanism bids on the behalf of all agent $i \notin \mathcal{B}_t$. Denote by b_{it} the bid of agent $i \in \mathcal{B}_t$ for the item at time t. The modification of $\mathcal{M}$ sets $b_{it} = \widehat{\mu}_{it}(t)$, for $i \notin B$. Then, the item is allocated at random to one of the agents in $\arg\max_i b_{it}$.

If i is the agent who received the item at time t, let $A = \{b_{jt} | j \in \mathcal{B}_t\} \cup \{\mu_{jt}|, j \notin \mathcal{B}_t\}$. Define γ_t as the second highest value in A. Let $\widehat{\gamma}_t(T)$ to be equal to $\max_{j\neq i} b_{jk}$. The payment of agent i will be

$$p_{it} \leftarrow \sum_{k=1}^{t-1} y_{ik} \min\{\widehat{\gamma}_k(t), b_{ik}\} - \sum_{k=1}^{t-1} p_{ik}.$$

To incorporate the fact that bidders can bid for an item, we must modify the definition of truthfulness.

DEFINITION 2. *Agent i is truthful if:*

1. $r_{it} = u_{it}$*, for all time* $x_{it} = 1, t \geq 1$.
2. *If i bids at time t, then* $E[|b_{it} - \mu_{it}|] \leq E[|\widehat{\mu}_{it} - \mu_{it}|]$.

Note that item 2 does not require that agent i bid their actual utility only that their bid be closer to the mark than the estimate. With this modification in definition, Theorems 1 and 3 continue to hold.

4. INDEPENDENT AND IDENTICALLY DISTRIBUTED UTILITIES

In this section, we assume that for each i, u_{it}'s are independent and identically-distributed random variables. For simplicity, we define $\mu_i = E[u_{it}], t > 0$. Without loss of generality, we also assume $0 < \mu_i \leq 1$.

In this environment, the learning algorithm we use is an ε-greedy algorithm for the multi-armed bandit problem[3]. Let $n_{it} = \sum_{k=1}^{t-1} x_{it}$. For $\epsilon \in (0,1)$, we define:

$$\begin{aligned} n_{it} &= \sum_{k=1}^{t-1} x_{it} \\ \eta_\epsilon(t) &= \min\{1, nt^{-\epsilon}\ln^{1+\epsilon} t\} \\ \widehat{\mu}_{it}(T) &= \begin{cases} (\sum_{k=1}^{T} x_{ik} r_{ik})/n_{iT}, & n_{iT} > 0 \\ 0, & n_{iT} = 0 \end{cases} \end{aligned}$$

Call the mechanism based on this learning algorithm $\mathcal{M}_\epsilon(iid)$.

LEMMA 4. *If all agents are truthful, then, under* $\mathcal{M}_\epsilon(iid)$

$$E[\Delta_t] = O(\frac{1}{\sqrt{t^{1-\epsilon}}}).$$

[3] See [3] for a similar algorithm.

The proof of this lemma is given in appendix A.

We show that $\mathcal{M}_\epsilon(iid)$, for $\varepsilon \le \frac{1}{3}$, satisfies all the desired properties we discussed in the previous section. Moreover, it satisfies a stronger notion of individual rationality. $\mathcal{M}_\epsilon(iid)$ satisfies *ex-post individual rationality* if for any agent i, and for all $T \ge 1$:

$$\sum_{t=1}^{T} p_{it} \le \sum_{t=1}^{T} x_{it} r_{it}$$

THEOREM 5. *$\mathcal{M}_\epsilon(iid)$ is ex-post individually rational. Also, for $0 \le \epsilon \le \frac{1}{3}$, $\mathcal{M}_\epsilon(iid)$ is asymptotically incentive compatible and ex-ante efficient.*

PROOF. We first prove ex-post individual rationality. It is sufficient to prove it only for the periods that agent i has received the item within an exploit phase. For T, such that $y_{iT} = 1$, we have:

$$\begin{aligned}\sum_{t=1}^{T} p_{it} &= \sum_{t=1}^{T-1} y_{it} \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\}\\ &\le \sum_{t=1}^{T-1} y_{it}\widehat{\gamma}_t(T) \le \sum_{t=1}^{T-1} y_{it}\widehat{\mu}_{iT}(T)\\ &\le n_{it}\widehat{\mu}_{iT}(T) = \sum_{t=1}^{T-1} x_{it} r_{it}\end{aligned}$$

The third inequality follows because the item is allocated to i at time T which implies $\widehat{\mu}_{iT}(T) \ge \widehat{\gamma}_t(T)$. We complete the proof by showing that conditions $(C1)$ and $(C2)$ hold. Note that $\mu_i \le 1$. By lemma 4, for $\epsilon \le \frac{1}{3}$:

$$E[1+\sum_{t=1}^{T-1} \Delta_t] = O(T^{\frac{1+\epsilon}{2}}) = o(T^{1-\epsilon} \ln^{1+\epsilon} T) = O(\sum_{t=1}^{T} \eta_\epsilon(t)\mu_i).$$

Therefore, $(C1)$ holds.

The welfare of any mechanism between time 1 and T is bounded by T. For any $\epsilon > 0$, $E[1 + \sum_{t=1}^{T-1} \Delta_t + \eta_t] = o(T)$ which implies $(C2)$. □

5. BROWNIAN MOTION

In this section, we assume for each i, $1 \le i \le n$, the evolution of μ_{it} is a reflected Brownian motion with mean zero and variance σ_i^2; the reflection barrier is 0. In addition, we assume $\mu_{i0} = 0$, and $\sigma_i^2 \le \sigma^2$, for some constant σ. The mechanism observes the values of μ_{it} at discrete times $t = 1, 2, \cdots$.

In this environment our learning algorithm estimates the reflected Brownian motion using a mean zero martingale. We define $\bar{l}_{it}$ is defined as the last time up to time t that the item is allocated to agent i. This includes both explore and exploit phases. If i has not been allocated any item yet, $\bar{l}_{it}$ is zero.

$$\eta_\epsilon(t) = \min\{1, nt^{-\epsilon} \ln^{2+2\epsilon} t\} \qquad (10)$$

$$\widehat{\mu}_{it}(T) = \begin{cases} r_{i\bar{l}_{it}} & t < T \\ r_{i\bar{l}_{i,t-1}} & t = T \\ r_{i\bar{l}_{i,T}} & t > T \end{cases} \qquad (11)$$

Call this mechanism $\mathcal{M}_\epsilon(\mathcal{B})$. For simplicity, we assume that the advertiser reports the exact value of μ_{it}. It is not difficult to verify that the results in this section hold as long as the expected value of the error of these estimates at time t is $o(t^{\frac{1}{6}})$.

We begin analyzing the mechanism by stating some well-known properties of reflected Brownian motions (see [7]).

PROPOSITION 6. *Let $[W_t, t \ge 0]$ be a reflected Brownian motion with mean zero and variance σ^2; the reflection barrier is 0. Assume the value of W_t at time t is equal to y:*

$$E[y] = \theta(\sqrt{t\sigma^2}) \qquad (12)$$

For $T > 0$, let $z = W_{t+T}$. For the probability density function of $z - y$ we have:

$$\begin{aligned}\Pr[(z-y) \in dx] &\le \sqrt{\frac{2}{\pi T\sigma^2}} e^{\frac{-x^2}{2T\sigma^2}} &(13)\\ \Pr[|z-y| \ge x] &\le \sqrt{\frac{8T\sigma^2}{\pi}} \frac{1}{x} e^{\frac{-x^2}{2T\sigma^2}} &(14)\\ E[|z-y| I(|z-y| \ge x)] &\le \sqrt{\frac{8T\sigma^2}{\pi}} e^{\frac{-x^2}{2T\sigma^2}} &(15)\end{aligned}$$

COROLLARY 7. *The expected value of the maximum of μ_{iT}, $1 \le i \le n$, is $\theta(\sqrt{T})$.*

Note that in the corollary above n and σ are constant. Now, similar to Lemma 4, we bound $E[\Delta_T]$. The proof is given in appendix B.

LEMMA 8. *Suppose under $\mathcal{M}_\epsilon(\mathcal{B})$ all agents are truthful until time T, then, $E[\Delta_T] = O(T^{\frac{\epsilon}{2}})$.*

Now we are ready to prove the main theorem of this section:

THEOREM 9. *$\mathcal{M}_\epsilon(\mathcal{B})$ is ex-post individually rational. Also, for $0 \le \epsilon \le \frac{1}{3}$, $\mathcal{M}_\epsilon(\mathcal{B})$ is asymptotically incentive compatible and ex-ante efficient.*

PROOF. We first prove ex-post individual rationality. It is sufficient to prove it only for the periods that agent i has received the item within an exploit phase. For T, such that $y_{iT} = 1$, we have:

$$\begin{aligned}\sum_{t=1}^{T} p_{it} &= \sum_{t=1}^{T-1} y_{it} \min\{\widehat{\gamma}_t(T), \widehat{\mu}_{it}(t)\}\\ &\le \sum_{t=1}^{T-1} y_{it}\widehat{\mu}_{it}(t) = \sum_{t=1}^{T-1} y_{it} r_{i\bar{l}_{i,t-1}}\\ &\le \sum_{t=1}^{T} x_{it} r_{it.}\end{aligned}$$

We complete the proof by showing the conditions $(C1)$ and $(C2)$ hold. By (12), the expected utility of each agent at time t from random exploration is

$$\theta(\sqrt{t\sigma^2} t^{-\epsilon} \ln^{1+\epsilon} t) = \theta(t^{\frac{1}{2}-\epsilon} \ln^{1+\epsilon} t).$$

Therefore, the expected utility up to time T from exploration is $\theta(T^{\frac{3}{2}-\epsilon} \ln^{1+\epsilon} T)$. By Lemma (8) and Corollary 7:

$$E[\max_{t \le T}\{\mu_{iT}\} + \sum_{t=1}^{T-1} \Delta_t] = O(T^{1+\frac{\epsilon}{2}}).$$

For $\epsilon \le \frac{1}{3}$, $\frac{3}{2} - \epsilon \ge 1 + \frac{\epsilon}{2}$ this yields Condition$(C1)$.

By Corollary 7, the expected value of $\max_i\{\mu_{iT}\}$ and γ_T are $\theta(\sqrt{T})$. Therefore, the expected welfare of an efficient mechanism between time 1 and T is $\theta(T^{\frac{3}{2}})$. For any $0 < \epsilon < 1$, we have:

$$\theta(T^{\frac{3}{2}}) = \omega(T^{\frac{3}{2}-\epsilon}\ln^{1+\epsilon} t + T^{1+\frac{\epsilon}{2}})$$

By condition $(C2)$, $\mathcal{M}_\epsilon(\mathcal{B})$ is asymptotically ex-ante efficient.

□

To apply this model to sponsored search we treat each item as a bundle of search queries. Each time step is defined by the arrival of m queries. The mechanism allocates all m queries to an advertiser and after that, the advertiser reports the average utility for these queries. The payment p_{it} is now the price per item, i.e. the advertiser pays mp_{it} for the bundle of queries. The value of m is chosen such that μ_{it} can be estimated with high accuracy.

6. DISCUSSION AND OPEN PROBLEMS

In this section we discuss some extensions of the mechanisms.

Multiple Slots. To modify $\mathcal{M}$ so that it can accommodate multiple slots we borrow from Gonen and Pavlov [12], who assume there exist a set of conditional distributions which determine the conditional probability that the ad in slot j_1 is clicked conditional on the ad in slot j_2 being clicked. During the exploit phase, $\mathcal{M}$ allocates the slots to the advertisers with the highest expected utility, and the prices are determined according to Holmstrom's lemma ([19], see also [1]) The estimates of the utilities are updated based on the reports, using the conditional distribution.

Delayed Reports. In some applications, the value of receiving the item is realized at some later date. For example, a user clicks on an ad and visits the website of the advertiser. A couple of days later, she returns to the website and completes a transaction. It is not difficult to adjust the mechanism to accommodate this setting by allowing the advertiser to report with a delay or change her report later.

Creating Multiple Identities. When a new advertiser joins the system, in order to learn her utility value our mechanism gives it a few items for free in the explore phase. Therefore our mechanism is vulnerable to advertisers who can create several identities and join the system.

It is not clear whether creating a new identity is cheap in our context because the traffic generated by advertising should eventually be routed to a legitimate business. Still, one way to avoid this problem is to charge users without a reliable history using CPC.

Acknowledgment. We would like to thank Arash Asadpour, Peter Glynn, Ashish Goel, Ramesh Johari, and Thomas Weber for fruitful discussions. The second author acknowledges the support from NSF and a gift from Google.

7. REFERENCES

[1] G. Aggarwal, A. Goel, and R. Motwani. Truthful auctions for pricing search keywords. *Proceedings of ACM conference on Electronic Commerce*, 2006.

[2] S. Athey, and I. Segal. An Efficient Dynamic Mechanism. *manuscript*, 2007.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning archive*, Volume 47 , Issue 2-3, 235-256, 2002.

[4] A. Bapna, and T. Weber. Efficient Dynamic Allocation with Uncertain Valuations. *Working Paper*, 2006.

[5] M. Balcan, A. Blum, J. Hartline, and Y. Mansour. Mechanism Design via Machine Learning. *Proceedings of 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[6] D. Bergemann, and J. Välimäki. Efficient Dynamic Auctions. *Proceedings of Third Workshop on Sponsored Search Auctions*, 2007.

[7] A. Borodin, and P. Salminen. Handbook of Brownian Motion: Facts and Formulae. *Springer*, 2002.

[8] A. Blum, V. Kumar, A. Rudra, and F. Wu. Online Learning in Online Auctions. *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete Algorithms*, 2003.

[9] R. Cavallo, D. Parkes, and S. Singh, Efficient Online Mechanism for Persistent, Periodically Inaccessible Self-Interested Agents. Working Paper, 2007.

[10] K. Crawford. Google CFO: Fraud A Big Threat. *CNN/Money*, December 2, 2004.

[11] J. Gittins. Multi-Armed Bandit Allocation Indices. *Wiley*, New York, NY, 1989.

[12] R. Gonen, and E. Pavlov. An Incentive-Compatible Multi-Armed Bandit Mechanism. *Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing*, 2007.

[13] B. Grow, B. Elgin, and M. Herbst. Click Fraud: The dark side of online advertising. *BusinessWeek*. Cover Story, October 2, 2006.

[14] N. Immorlica, K. Jain, M. Mahdian, and K. Talwar. Click Fraud Resistant Methods for Learning Click-Through Rates. *Proceedings of the 1st Workshop on Internet and Network Economics*, 2005.

[15] B. Kitts, P. Laxminarayan, B. LeBlanc, and R. Meech. A Formal Analysis of Search Auctions Including Predictions on Click Fraud and Bidding Tactics. *Workshop on Sponsored Search Auctions*, 2005.

[16] R. Kleinberg. Online Decision Problems With Large Strategy Sets. *Ph.D. Thesis*, MIT, 2005.

[17] S. Lahaie, and D. Parkes. Applying Learning Algorithms to Preference Elicitation. *Proceedings of the 5th ACM conference on Electronic Commerce*, 2004.

[18] M. Mahdian, and K. Tomak. Pay-per-action model for online advertising. *Proceedings of the 3rd International Workshop on Internet and Network Economics*, 549-557, 2007.

[19] P. Milgrom, Putting Auction Theory to Work. *Cambridge University Press*, 2004.

[20] D. Mitchell. Click Fraud and Halli-bloggers. *New York Times*, July 16, 2005.

[21] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors. Algorithmic Game Theory, *Cambridge University Press*, 2007.

[22] D. Parkes. Online Mechanisms *Algorithmic Game Theory (Nisan et al. eds.)*, 2007.

[23] B. Stone. When Mice Attack: Internet Scammers Steal Money with "Click Fraud". *Newsweek*, January 24, 2005.

[24] R. Wilson. Game-Theoretic Approaches to Trading Processes. *Economic Theory: Fifth World Congress*, ed. by T. Bewley, chap. 2, pp. 33-77, Cambridge University Press, Cambridge, 1987.

[25] J. Wortman, Y. Vorobeychik, L. Li, and J. Langford. Maintaining Equilibria During Exploration in Sponsored Search Auctions. *Proceedings of the 3rd International Workshop on Internet and Network Economics*, 2007.

APPENDIX

A. PROOF OF LEMMA 4

Proof. We prove the lemma by showing that for any agent i,

$$\Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i] = o(\frac{1}{t^c}), \forall c > 0.$$

First, we estimate $E[n_{it}]$. There exists a constant d such that:

$$E[n_{it}] \geq \sum_{k=1}^{t-1} \frac{\eta_\epsilon(k)}{n} = \sum_{k=1}^{t-1} \min\{\frac{1}{n}, k^{-\epsilon} \ln^{1+\epsilon} k\} > \frac{1}{d} t^{1-\epsilon} \ln^{1+\epsilon} t$$

By the Chernoff-Hoeffding bound:

$$\Pr[n_{it} \leq \frac{E[n_{it}]}{2}] \leq e^{\frac{-t^{1-\epsilon} \ln^{1+\epsilon} t}{8d}}.$$

Inequality (1) and the Chernoff-Hoeffding bound imply:

$$\begin{aligned}
\Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i] &= \Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i \wedge n_{it} \geq \frac{E[n_{it}]}{2}] \\
&\quad + \Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i \wedge n_{it} < \frac{E[n_{it}]}{2}] \\
&\leq 2e^{\frac{-\frac{1}{t^{1-\epsilon}} t^{1-\epsilon} \ln^{1+\epsilon} t \; \mu_i}{2d}} + e^{\frac{-t^{1-\epsilon} \ln^{1+\epsilon} t}{8d}} \\
&= o(\frac{1}{t^c}), \forall c > 0
\end{aligned}$$

Therefore, with probability $1 - o(\frac{1}{t})$, for all agents, $\Delta_t \leq \frac{1}{\sqrt{t^{1-\epsilon}}}$. Since the maximum value of u_{it} is 1, $E[\Delta_t] = O(\frac{1}{\sqrt{t^{1-\epsilon}}})$. □

B. PROOF OF LEMMA 8

Proof. Define $X_{it} = |\mu_{i,T} - \mu_{i,T-t}|$. We first prove $\Pr[X_{it} > T^{\frac{\epsilon}{2}}] = o(\frac{1}{T^c}), \forall c > 0$. There exists a constant T_d such that for any time $T \geq T_d$, the probability that i has not been randomly allocated the item in the last $t < T_d$ step is at most:

$$\Pr[T - \bar{l}_{i,T-1} > t] < (1 - T^{-\epsilon} \ln^{2+2\epsilon} T)^t \leq e^{\frac{-t \ln^{2+2\epsilon} T}{T^\epsilon}}. \quad (16)$$

Let $t = \frac{1}{\ln^{1+\epsilon} T} T^\epsilon$. By equation (14) and (16),

$$\begin{aligned}
\Pr[X_{it} > T^{\frac{\epsilon}{2}}] &= \Pr[X_{it} > T^{\frac{\epsilon}{2}} \wedge T - \bar{l}_{i,T-1} \leq t] \\
&\quad + \Pr[X_{it} > T^{\frac{\epsilon}{2}} \wedge T - \bar{l}_{i,T-1} > t] \\
&= o(\frac{1}{T^c}), \forall c > 0.
\end{aligned}$$

Hence, with high probability, for all the n agents, $X_{it} \leq T^{\frac{\epsilon}{2}}$. If for some of the agents $X_{it} \geq T^{\frac{\epsilon}{2}}$, then, by Corollary 7, the expected value of the maximum of μ_{it} over these agents is $\theta(\sqrt{T})$. Therefore, $E[\max_i\{X_{it}\}] = O(T^{\frac{\epsilon}{2}})$. The lemma follows because $E[\Delta_T] \leq E[\max_i\{X_{it}\}]$. □