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

Hamid Nazerzadeh[*]   Amin Saberi[*]   Rakesh Vohra[†]

September 9, 2008

Abstract

We examine the problem of allocating a resource repeatedly over time amongst a set of agents. The utility that each agent derives from consumption of the item is private information to that agent, and prior to consumption, may be unknown to that agent. 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.

Our mechanism can be interpreted as a 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.

---

[*]Management Science and Engineering Department, Stanford University, Stanford, CA 94305, email: {hamidnz,saberi@stanford.edu}

[†]Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, email: {r-vohra@kellogg.northwestern.edu}

1 Introduction

A collection of self-interested agents compete 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 realizes 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] [1].

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. The mechanism takes two randomly *alternating* actions: exploration and exploitation. During exploration, the item is allocated for free to a randomly chosen agent. During exploitation, the mechanism allocates the item to the agent with the highest estimated expected utility. After each allocation, the agent who has received the item reports its realized utility. 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 informed agents to bid directly.

In the next section, we motivate our work in the context of online advertising. However, the motivation for our mechanism is not limited to such applications.

---

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

1.1 Applications to Online Advertising

Currently, the two main charging models in the online advertising industry are cost-per-impression (CPM) and cost-per-click (CPC). In the CPM, 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. The last few years have seen a shift towards CPC. CPC has been 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 focus on a recently popular and widely advocated charging scheme known as the Cost-Per-Action or Cost-Per-Acquisition (CPA) 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 briefly describe a few advantages of CPA over CPC and refer the reader to [18] for a more detailed discussion.

One drawback of CPC 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 [14, 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.[2]

A fundamental difference between CPA and CPC is that a click on an advertisement can be observed by both advertiser and publisher. However, the action of a user is hidden from the publisher and observable only by the advertiser. Although a publisher can require advertisers to install 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. It is this difference that motivates the present paper.

In our setting, the item being allocated is a search query for a keyword. An advertiser obtains a payoff when the user clicks on her advertisement and takes a specific action. Since the payoff is uncertain, she cannot know what it will be unless her ad is displayed. For simplicity of exposition only, we assume one keyword and one advertisement slot. In section **??** we outline how to extend our results to the case where more than one advertisement can be displayed for each query.

---

[2]Of course in this case, CPA makes generating a fraudulent action a more costly enterprize, but not impossible (one could use a stolen credit card number for example.).

1.2 Related Work

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

Most of these works, like [5, 8, 11, 17], consider one-shot games or repeated auctions in which the agents depart upon receiving an item. In our setting we may allocate the 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 allocate 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 that may 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 [12]). 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 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 [13] propose a CPC, learning based auction when there is uncertainty about click through rates 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 paper, unlike ours, the distribution of the utilities of agents are assumed fixed over time.

Immorlica et al. [15], 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 times $t = 1, 2, \cdots$. 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

3

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.

Let $\mathcal{M}$ be a mechanism. At each time, $\mathcal{M}$ allocates the item to one of the agents. Define x_{it} to be the variable indicating the allocation of the item to i at time t. If the item is allocated to agent i, she can observe u_{it}, her utility for the item. Then she reports r_{it} as her utility to the mechanism. The mechanism then determines the payment, denoted by p_{it}. Note that we do not require an agent to know her utility for the item before acquiring it. She may also misreport her utility to the mechanism after the allocation.

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:

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

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 and Revenue: Call a mechanism ex-ante efficient if at each time t it allocates the item to an agent in $\operatorname{argmax}_i\{\mu_{it}\}$. 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 γ_t be the second highest μ_{it} at time $t > 0$. Then, the expected revenue of a second price mechanism up to time T is equal to $E[\sum_{t=1}^{T} \gamma_t]$.

We will measure the efficiency and the revenue of $\mathcal{M}$ by comparing it to the second price mechanism that allocates the item to an agent in $\operatorname{argmax}_i\{\mu_{it}\}$ and charges her the second highest μ_{it}. Let $W(T)$ and $R(T)$ be the expected welfare and the expected revenue of mechanism $\mathcal{M}$ between time 1 and T, when all agents are truthful, i.e.

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

$$R(T) = E[\sum_{t=1}^{T} \sum_{i=1}^{n} p_{it}]$$

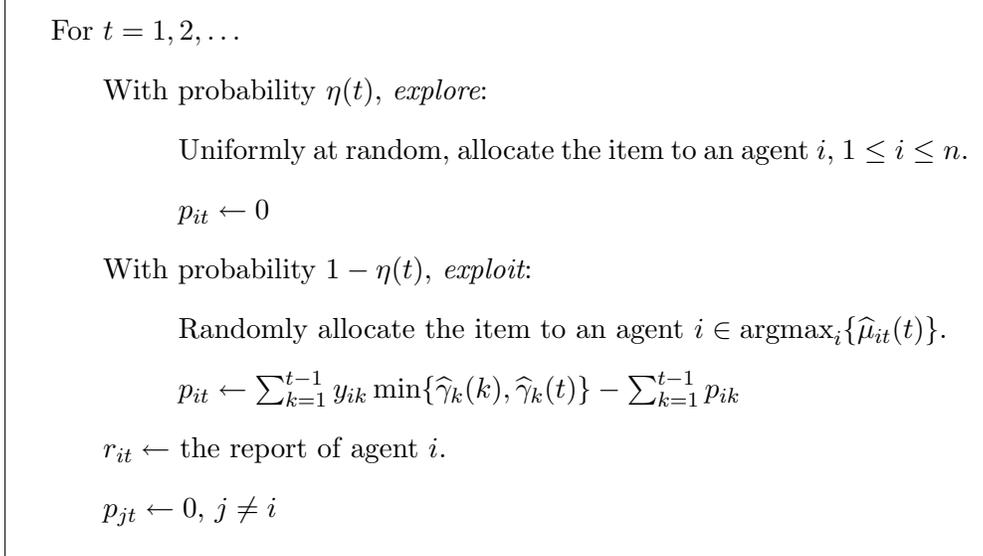For $t = 1, 2, \ldots$

    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 $i \in \mathrm{argmax}_i \{\widehat{\mu}_{it}(t)\}$.

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

    $r_{it} \leftarrow$ the report of agent i.

    $p_{jt} \leftarrow 0$, $j \neq i$

Figure 1: Mechanism $\mathcal{M}$

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

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

Also, the revenue of $\mathcal{M}$ is asymptotically equivalent to the revenue of the second price auction if:

$$R(T) = (1 - o(1))E[\sum_{t=1}^{T} \gamma_t]$$

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 4.1). In section 5.1 and 5.2 we give two examples of environments where learning algorithms satisfying these sufficient conditions exist.

    The mechanism randomly alternates between two actions: *exploration* and *exploitation*. At time t, with probability $\eta(t)$, $\eta : \mathbb{N} \to [0, 1]$, the mechanism explores i.e. it allocates the item for free to an agent chosen uniformly at random. With the remaining probability, the mechanism exploits. During exploitation, the item is allocated to the agent with the highest estimated expected utility. Then, 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 uses the history of the reports of agent i to give an estimate of μ_{it}. Let $\widehat{\mu}_{it}(T)$ be the estimate of the learning algorithm for μ_{it} conditional on the history of the reports

up to but not including time T. Note that T can be bigger than t. In other words, we allow the learning algorithm to refine earlier estimates using more recent history.

We now describe the payment scheme. Let $\widehat{\gamma}_t(T) = \max_{j \neq i}\{\widehat{\mu}_{jt}(T)\}$, where i is the agent who receives the item at time t. We define y_{it} to be the indicator variable of the allocation of the item to agent i during exploitation. The payment of agent i at time t, denoted p_{it}, is equal to:

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

Therefore, we have:

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

In this payment scheme, an agent only pays for the items that are allocated to her during exploitation, up to but not including time t. The payments emulate the second pricing scheme with the difference that the second highest utility γ_t is replaced with its estimate.

Another important feature of our payment mechanism is that it is adaptive and cumulative. We allow the mechanism to adjust the payments of the agents if the items allocated to them were overpriced earlier.

4 Analysis

Assume all agents are truthful up to time t. Let Δ_t be the maximum over all agents i, of the difference between μ_{it} and its estimate using only reports taken during exploration. We assume the accuracy of the learning algorithm is monotone, i.e., at time $T \geq t$ we have:

$$E[|\widehat{\mu}_{it}(T) - \mu_{it}|] \leq E[|\widehat{\mu}_{it}(t) - \mu_{it}|] \leq E[\Delta_t] \tag{1}$$

In the inequality 1, and in the rest of the paper, the expectations of $\widehat{\mu}_{it}$ are condition on 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.

In this section, we will relate the performance of the mechanism to the estimation error of the learning algorithm. We show first that if $\sum \Delta_t$ is small, then agents cannot gain much by deviating from the truthful strategy. We also bound the efficiency loss in terms of $\sum \Delta_t$.

Theorem 1 *For a truthful agent i up to time T, the expected amount that i may be overcharged for the items she receives is bounded by the total estimation error of the algorithm, i.e.,*

$$E[\sum_{t=1}^{T} y_{it} u_{it}] - E[\sum_{t=1}^{T} p_{it}] \geq -E[\sum_{t=1}^{T} \Delta_t].$$

Proof :   We prove a stronger result by showing:

$$E[\sum_{t=1}^{T} p_{it}] - E[\sum_{t=1}^{T} y_{it} u_{it}] \leq E[\sum_{t=1}^{T-1} y_{it}(\widehat{\gamma}_t(t) - \mu_{it})^+] \leq E[\sum_{t=1}^{T} \Delta_t] \tag{2}$$

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

If $y_{it} = 0$ then $p_{it} = 0$. Also, recall that the expectations are taken conditioned on the previous history. Therefore, for every time t, $E[u_{it}] = E[\mu_{it}]$. By the payment rule we have:

$$
\begin{aligned}
E[\sum_{t=1}^{T} p_{it}] - E[\sum_{t=1}^{T} y_{it} u_{it}] &\leq E[\sum_{t=1}^{T-1} y_{it}(\min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\} - \mu_{it})] \\
&\leq E[\sum_{t=1}^{T-1} y_{it}(\widehat{\gamma}_t(t) - \mu_{it})] \\
&\leq E[\sum_{t=1}^{T-1} (\widehat{\gamma}_t(t) - \mu_{it})^+]
\end{aligned}
$$

y_{it} indicates whether the mechanism allocated the item to agent i at time t. Therefore, $y_{it} = 1$ implies $\widehat{\gamma}_t(t) \leq \widehat{\mu}_{it}(t)$. Plugging that into the above inequality, we get:

$$
\begin{aligned}
E[\sum_{t=1}^{T} p_{it}] - E[\sum_{t=1}^{T} y_{it} u_{it}] &\leq E[\sum_{t=1}^{T-1} (\widehat{\mu}_{it}(t) - \mu_{it})^+] \\
&\leq E[\sum_{t=1}^{T-1} \Delta_t]
\end{aligned}
$$

The last inequality follows from equation (1). $\qquad\square$

Now we study the incentive compatibility aspects of the mechanism.

Theorem 2 *If all other agents are truthful, agent i cannot increase her expected utility up to time T by more than the quantity below if she deviates from the truthful strategy:*

$$
8E[\sum_{t=1}^{T} \Delta_t] + E[\max_{1 \leq t \leq T}\{\mu_{it}\}] \tag{3}
$$

Proof : Let us bound the expected profit of i for deviating from the truthful strategy. The term $E[\max_{1 \leq t \leq T}\{\mu_{it}\}]$ in the expression above bounds the outstanding payment of agent i up to time T. Recall that agents do not pay for the last item they receive during exploitation.

Let $\mathcal{S}$ be the strategy that i deviates to. Fixing the evolution of all u_{jt}'s, $1 \leq j \leq n$, and all random choices of the mechanism. Let D_T be the set of times that i receives the item under strategy $\mathcal{S}$ during exploitation and before time T. Formally, $D_T = \{t < T | y_{it} = 1,$ if the strategy of i is $\mathcal{S}\}$. Similarly, let $C_T = \{t < T | y_{it} = 1,$ if i is truthful$\}$. Also, let $\widehat{\mu}'_{it}$, and $\widehat{\gamma}'_t$ correspond to the estimates of the mechanism when the strategy of i is $\mathcal{S}$. The expected profit i could obtain under strategy S from the items she received during exploitation, up to time $T - 1$, is equal to:

$$
\begin{aligned}
E[\sum_{t \in D_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}] &= E[\sum_{t \in D_T \cap C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}] + \\
&\quad E[\sum_{t \in D_T \setminus C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}] \tag{4}
\end{aligned}
$$

For times $t \geq 1$, we examine two cases:

7

1. The first case is when $t \in D_T \cap C_T$. We show that in this case the "current price" given to agent i under the two scenarios are close. To this aim, we first observe:

$$\min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} = \widehat{\gamma}_t(t) + \min\{\widehat{\gamma}'_t(t) - \widehat{\gamma}_t(t), \widehat{\gamma}'_t(T) - \widehat{\gamma}_t(t)\}$$

Let $j \neq i$ be the agent with the highest $\widehat{\mu}_{jt}(t)$, i.e., $\widehat{\mu}_{jt}(t) = \text{argmax}_{j \neq i}\{\widehat{\mu}_{jt}(t)\} = \widehat{\gamma}_t(t)$. Because i is the winner both in D_T and C_T, by definition of $\widehat{\gamma}_t$ we have $\widehat{\mu}'_{jt}(t) \leq \widehat{\gamma}'_t(t)$ and $\widehat{\mu}'_{jt}(T) \leq \widehat{\gamma}'_t(T)$. Plugging into the above inequality we get:

$$
\begin{aligned}
\min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} &\geq \widehat{\gamma}_t(t) + \min\{\widehat{\mu}'_{jt}(t) - \widehat{\mu}_{jt}(t), \widehat{\mu}'_{jt}(T) - \widehat{\mu}_{jt}(t)\} \\
&\geq \widehat{\gamma}_t(t) - |\widehat{\mu}_{jt}(t) - \widehat{\mu}'_{jt}(t)| - |\widehat{\mu}_{jt}(t) - \widehat{\mu}'_{jt}(T)| \\
&= \widehat{\gamma}_t(t) - |\widehat{\mu}_{jt}(t) - \mu_{jt} + \mu_{jt} - \widehat{\mu}'_{jt}(t)| - |\widehat{\mu}_{jt}(t) - \mu_{jt} + \mu_{jt} - \widehat{\mu}'_{jt}(T)| \\
&\geq \widehat{\gamma}_t(t) - 2|\widehat{\mu}_{jt}(t) - \mu_{jt}| - |\mu_{jt} - \widehat{\mu}'_{jt}(t)| - |\mu_{jt} - \widehat{\mu}'_{jt}(T)|
\end{aligned}
$$

By taking expectation from both sides we have:

$$
\begin{aligned}
E[\min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}I(t \in D_T \cap C_T)] \geq\ & E[\widehat{\gamma}_t(t)I(t \in D_T \cap C_T)] - \\
& E[2|\widehat{\mu}_{jt}(t) - \mu_{jt}|I(t \in D_T \cap C_T)] - \\
& E[(|\mu_{jt} - \widehat{\mu}'_{jt}(t)| - |\mu_{jt} - \widehat{\mu}'_{jt}(T)|)\,I(t \in D_T \cap C_T)] \\
\geq\ & E[\widehat{\gamma}_t(t)I(t \in D_T \cap C_T)] - \\
& E[2|\widehat{\mu}_{jt}(t) - \mu_{jt}| + |\mu_{jt} - \widehat{\mu}'_{jt}(t)| + |\mu_{jt} - \widehat{\mu}'_{jt}(T)|]
\end{aligned}
$$

Because agent j is truthful, by (1), we get:

$$E[\min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}I(t \in D_T \cap C_T)] \geq E[\widehat{\gamma}_t(t)I(t \in D_T \cap C_T)] - 4E[\Delta_t] \qquad (5)$$

2. The second case is when $t \in D_T \setminus C_T$. We show in this case agent i cannot increase her "profit" by much.

Let j be the agent who would receive the item at time t when agent i is truthful. Therefore $\widehat{\mu}_{jt}(t) \geq \widehat{\mu}_{it}(t)$. Also, $\widehat{\mu}_{jt}(t) \leq \max_{j \neq i}\{\widehat{\mu}_{jt}\} = \gamma_t(t)$. Similarly, $\widehat{\mu}_{jt}(T) \leq \gamma_t(T)$.

$$
\begin{aligned}
\mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} &= \widehat{\mu}_{it}(t) + |\mu_{it} - \widehat{\mu}_{it}(t)| - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} \\
&\leq \widehat{\mu}_{jt}(t) + |\mu_{it} - \widehat{\mu}_{it}(t)| - \min\{\widehat{\mu}_{jt}(t), \widehat{\mu}_{jt}(T)\}] \\
&= \max\{0, \widehat{\mu}_{jt}(t) - \widehat{\mu}_{jt}(T)\} + |\mu_{it} - \widehat{\mu}_{it}(t)| \\
&\leq |\widehat{\mu}_{jt}(t) - \widehat{\mu}_{jt}(T)| + |\mu_{it} - \widehat{\mu}_{it}(t)| \\
&\leq |\widehat{\mu}_{jt}(t) - \mu_{jt}| + |\mu_{jt} - \widehat{\mu}_{jt}(T)| + |\mu_{it} - \widehat{\mu}_{it}(t)|
\end{aligned}
$$

Sum up the inequality above over all $t \in D_T \setminus C_T$. Because all agents are truthful, taking expectation from both sides, , by (1), we get:

$$E[\sum_{t \in D_T \setminus C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}] \leq 3E[\sum_{t=1}^{T} \Delta_t] \qquad (6)$$

8

Substituting inequalities (5) and (6) into (4):

$$E[\sum_{t=1}^{T-1} y_{it}u_{it} - p_{it}] \leq E[\sum_{t \in D_T \cap C_T} \mu_{it} - \widehat{\gamma}_t(t)] + 7E[\sum_{t=1}^{T} \Delta_t]$$

$$= E[\sum_{t \in C_T} \mu_{it} - \widehat{\gamma}_t(t)] - E[\sum_{t \in C_T \setminus D_T} \mu_{it} - \widehat{\gamma}_t(t)] + 7E[\sum_{t=1}^{T} \Delta_t]$$

By inequality (2), $-E[\sum_{t \in C_T \setminus D_T} \mu_{it} - \gamma_t] \leq E[\sum_{t=1}^{T} \Delta_t]$. Therefore:

$$E[\sum_{t=1}^{T-1} y_{it}u_{it} - p_{it}] - E[\sum_{t \in C_T} u_{it} - p_{it}] \leq 8E[\sum_{t=1}^{T} \Delta_t]$$

which completes the proof. $\square$

We now compare the welfare of our mechanism to the efficient mechanism that allocates the item to the agent with the highest expected utility every time. The expected loss of efficiency during exploration is equal to $E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]$. In the next theorem, we show that in the equilibrium, the efficiency loss during exploitation is bounded by a factor of the total estimation error of the learning algorithm.

Theorem 3 *Let $W(T)$ denote the expected welfare of mechanism $\mathcal{M}$ between time 1 and T. If all the agents are truthful, we have:*

$$W(T) \geq E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\} + 2\Delta_t]$$

Proof : Our mechanism can lose efficiency in two ways. First, during exploration when we allocate the item to an agent chosen uniformly at random. The expected loss in this case is at most $E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]$.

Second, during exploitation, the error in the estimates may lead to allocating the item to an agent who does not value the item the most. Suppose at time t, during exploitation, the mechanism allocated the item to agent j instead of i, when $\mu_{jt} < \mu_{it}$. By the rule of the mechanism we have $\widehat{\mu}_{it}(t) \leq \widehat{\mu}_{jt}(t)$. By subtracting this inequality from $\mu_{jt} - \mu_{it}$ we get:

$$\mu_{jt} - \mu_{it} \geq \mu_{jt} - \mu_{it} - (\widehat{\mu}_{jt}(t) - \widehat{\mu}_{it})$$
$$= (\mu_{jt} - \widehat{\mu}_{jt}(t)) + (\widehat{\mu}_{it}(t) - \mu_{it})$$

We sum up this inequality over all such time t, and by inequality (1), the expected efficiency loss during exploration is bounded by $2E[\sum_{t=1}^{T} \Delta_t]$.

Therefore, for the expected welfare of $\mathcal{M}$ between time 1 and T we have:

$$E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - W(T) \leq E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\} + 2\Delta_t]$$

$\square$

4.1 Sufficient Conditions for the Learning Algorithm

In this section, we give sufficient conditions on the learning algorithm which guarantee asymptotic ex-ante individual rationality, incentive compatibility and efficiency of our mechanism.

Theorem 4 *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 asymptotically incentive compatible. Also, if in addition to $(C1)$, the following condition holds

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

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

Theorem 4 suggests a trade-off between exploitation and exploration rates in our context: higher exploration rates lead to more accurate estimates of the utilities of the agents but they decrease the efficiency. So it is natural that condition $(C1)$ gives a *lower bound* on the exploration rate and condition $(C2)$ provides an *upper bound*. In the following sections, we 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 asymptotic efficiency and incentive compatibility.

Proof : The expected utility of a truthful agent i up to time T is equal to:

$$E[\sum_{t=1}^{T} x_{it}u_{it}] \;\; = \;\; \frac{1}{n}E[\sum_{t=1}^{T} \eta(t)\mu_{it}] + E[\sum_{t=1}^{T} y_{it}\mu_{it}]$$

Subtracting $E[\sum_{t=1}^{T} p_{it}]$ from both sides, by Theorem 1 we get:

$$E[\sum_{t=1}^{T} x_{it}u_{it}] - E[\sum_{t=1}^{T} p_{it}] \;\; = \;\; \frac{1}{n}E[\sum_{t=1}^{T} \eta(t)\mu_{it}] + (E[\sum_{t=1}^{T} y_{it}\mu_{it}] - E[\sum_{t=1}^{T} p_{it}])$$

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

Plugging condition $(C1)$ into the equation above yields:

$$E[\sum_{t=1}^{T} x_{it}u_{it}] - E[\sum_{t=1}^{T} p_{it}] \;\; \geq \;\; (\frac{1}{n} - o(1))E[\sum_{t=1}^{T} \eta(t)\mu_{it}] \tag{7}$$

Therefore, the mechanism is asymptotically ex-ante individually rational. Moreover, inequality (7) implies that the utility of the agent i is $\Omega(E[\sum_{t=1}^{T} \eta(t)\mu_{it}])$. Thus, by Theorem 2, if $(C1)$ holds, then the mechanism is asymptotically incentive compatible.

To prove the claim about the efficiency of the mechanism, we invoke Theorem 3. By this theorem and condition $(C1)$ we have:

$$W(T) \;\; \geq \;\; E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - E[\sum_{t=1}^{T} \eta(t)\max_i\{\mu_{it}\} + 2\Delta_t]$$

$$\geq \;\; E[\sum_{t=1}^{T} \max_i\{\mu_{it}\}] - (1 + o(1))E[\sum_{t=1}^{T} \eta(t)\max_i\{\mu_{it}\}]$$

Plugging condition $(C2)$ into the equation above we get:

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

which implies asymptotic ex-ante efficiency. $\qquad\square$

Theorem 4 shows 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 give similar conditions on the revenue guarantee of the mechanism.

Theorem 5 *If in addition to $(C1)$, the following condition holds*

$$(C3) \quad E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}] = o(E[\sum_{t=1}^{T} \gamma_{it}])$$

then the revenue of the mechanism is asymptotically equivalent to the revenue of the efficient mechanism that at every time allocates the item to the agent with the highest expected utility and charges the winning agent the second highest expected utility.

Proof : Using a similar argument as before, our mechanism can lose revenue in three ways. The first one is the loss during the exploration which is at most $E[\sum_{t=1}^{T} \eta(t)\gamma_t] \leq E[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}]$. There is also an estimation error of γ_t. Let i be the agent who has received the item at time t. We consider two cases:

1. If i is the agent with the second highest expected utility, then let j be the agent with the highest expected utility. The estimation error of γ_t is equal to $\gamma_t - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\}$.

$$
\begin{aligned}
\gamma_t - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\} &\leq \mu_{it} - \min\{\widehat{\mu}_{jt}(t) - \widehat{\mu}_{jt}(T)\} \\
&\leq \mu_{jt} - \min\{\widehat{\mu}_{jt}(t) - \widehat{\mu}_{jt}(T)\} \\
&\leq \max\{\mu_{jt} - \widehat{\mu}_{jt}(t), \mu_{jt} - \widehat{\mu}_{jt}(T)\} \\
&\leq |\mu_{jt} - \widehat{\mu}_{jt}(t)| + |\mu_{jt} - \widehat{\mu}_{jt}(T)|
\end{aligned}
$$

Therefore in this case, by inequality (1), the expected estimation error of γ_t is bounded by $2\Delta_t$.

2. Otherwise, let j be the agent with the second highest expected utility.

$$\gamma_t - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\} \leq \mu_{jt} - \min\{\widehat{\mu}_{jt}(t) - \widehat{\mu}_{jt}(T)\}$$

Similar to the previous case, we have:

$$\gamma_t - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\} \leq |\mu_{jt} - \widehat{\mu}_{jt}(t)| + |\mu_{jt} - \widehat{\mu}_{jt}(T)|$$

which bounds the expected estimation error of γ_t by $2\Delta_t$.

11

The third factor contributing to loss of revenue is the outstanding payment of the agents. Agents do not pay for the last item they have received during exploitation. These outstanding payments contribute to a loss that is bounded by $n \cdot E[\max_{1 \leq t \leq T} \gamma_t]$.

For the expected revenue of the mechanism we have:

$$
\begin{aligned}
R(T) &= E[\sum_{t=1}^{T} \sum_{i=1}^{n} p_{it}] \\
&\geq E[\sum_{t=1}^{T}(1 - \eta(t)) \min\{\widehat{\gamma}_t(T), \widehat{\gamma}_t(t)\}] - n \cdot E[\max_{t \leq T} \gamma_t] \\
&\geq E[\sum_{t=1}^{T}(1 - \eta(t))(\gamma_t - 4\Delta_t)] - n \cdot E[\max_{t \leq T} \gamma_t] \\
&\geq E[\sum_{t=1}^{T} \gamma_t] - E[\sum_{t=1}^{T} \eta(t)\gamma_t] - E[\sum_{t=1}^{T} 4\Delta_t] - n \cdot E[\max_{t \leq T} \gamma_t]
\end{aligned}
$$

Plugging condition $(C1)$ we get:

$$
R(T) \geq E[\sum_{t=1}^{T} \gamma_t] - (1 + o(1))E[\sum_{t=1}^{T} \eta(t) \max_i \{\mu_{it}\}]
$$

Therefore, by condition $(C3)$

$$
E[\sum_{t=1}^{T} \gamma_t] - R(T) = o(E[\sum_{t=1}^{T} \gamma_t]).
$$

$\square$

4.2 Allowing the 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 the learning algorithm computes from them. This may be advantageous when the bidders themselves are unaware of their expected utilities. However, sometimes the agents have a better estimate of their utilities than the mechanism. For this reason we describe how to modify $\mathcal{M}$ so as to allow the agents to bid for the items.

Suppose $\mathcal{M}$ is doing exploitation at time t and let $\mathcal{B}_t$ be the set of agents who are bidding 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 $\text{argmax}_i b_{it}$.

Let i be the agent who received the item at time t. Also, let $\widehat{\gamma}_t(T)$ to be equal to $\max_{j \neq i}\{b_{jk}\}$. The payment of agent i will be

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

We also call agent i truthful if:

1. $r_{it} = u_{it}$, for all time $x_{it} = 1, t \geq 1$.

2. If i bids at time t, then $E[|b_{it} - \mu_{it}|] \leq E[|\widehat{\mu}_{it}(t) - \mu_{it}|]$.

Note that the second condition does not require that agents bid their actual utility, only that their bids are closer to their actual utilities than our estimates. With these modifications, the theorems in the previous section continue to hold.

5 Examples

In this section, we study two models for the evolution of utilities of the agents. In the first, the utilities of the agents are independent and identically-distributed. In the second, the utility of each agent evolves independently like a reflected Browning motion. In both instances, we give simple sampling-based algorithms for learning the utilities of the agents. We show how we can use Theorems 4 and 5 to adjust the exploration rate of a simple learning algorithm to satisfy the conditions of our theorem. We also show that these mechanisms satisfy a stronger notion of individual rationality i.e. *ex-post individual rationality*. A mechanism is ex-post individually rational if for any agent i and for all $T \geq 1$:

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

5.1 Independent and Identically-Distributed Utilities

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

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

The proof of this lemma is given in appendix A.1.

---

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

Theorem 7 $\mathcal{M}_\epsilon(iid)$ *is ex-post individually rational. Also, for* $0 \leq \epsilon \leq \frac{1}{3}$, $\mathcal{M}_\epsilon(iid)$ *is asymptotically incentive compatible, ex-ante efficient, and has revenue asymptotically equivalent to the revenue of the efficient second price auction.*

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 during exploitation. For T, such that $y_{iT} = 1$, we have

$$\sum_{t=1}^{T} p_{it} = \sum_{t=1}^{T-1} y_{it} \min\{\widehat{\gamma}_t(T), \widehat{\gamma}_{it}(t)\} \leq \sum_{t=1}^{T-1} y_{it} \widehat{\gamma}_t(T).$$

Because $y_{iT} = 1$ we have $\widehat{\gamma}_t(T) \leq \widehat{\mu}_{it}(T)$. Plugging into the inequality above we get

$$\sum_{t=1}^{T} p_{it} \leq \sum_{t=1}^{T-1} y_{it} \widehat{\mu}_{it}(T) \leq \sum_{t=1}^{T-1} x_{it} \widehat{\mu}_{it}(T) = \sum_{t=1}^{T-1} x_{it} r_{it}.$$

Therefore the mechanism is ex-post individually rational. We complete the proof by showing that conditions $(C1)$, $(C2)$, and $(C3)$ hold. Note that $\mu_i \leq 1$. By Lemma 6, for $\epsilon \leq \frac{1}{3}$, we have

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

Therefore, $(C1)$ holds.

The welfare and revenue of the mechanism between time 1 and T of is $\theta(T)$. For any $\epsilon > 0$, $E[1 + \sum_{t=1}^{T-1} \Delta_t + \eta_t] = o(T)$ which satisfies $(C2)$ and $(C3)$. $\square$

5.2 Brownian Motion

Here we assume that for each i, $1 \leq i \leq n$, 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 $\sigma_i^2 \leq \sigma^2$, for some constant σ.

In this environment our learning algorithm estimates the reflected Brownian motion using a mean zero martingale. We define l_{it} as the last time up to time t that the item is allocated to agent i. This includes both exploration and exploitation actions. If i has not been allocated any item yet, l_{it} is zero.

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

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

Call this mechanism $\mathcal{M}_\epsilon(\mathcal{B})$. 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 $o(t^{\frac{1}{6}})$. However, for simplicity of exposition, we assume that the advertiser reports the exact value of μ_{it}.

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

Proposition 8 *Let $[W_t, t \geq 0]$ 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:*

$$E[y] = \theta(\sqrt{t\sigma^2}) \tag{8}$$

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

$$\Pr[(z - y) \in dx] \leq \sqrt{\frac{2}{\pi T\sigma^2}} e^{\frac{-x^2}{2T\sigma^2}} \tag{9}$$

$$\Pr[|z - y| \geq x] \leq \sqrt{\frac{8T\sigma^2}{\pi}} \frac{1}{x} e^{\frac{-x^2}{2T\sigma^2}} \tag{10}$$

$$E[|z - y|I(|z - y| \geq x)] \leq \sqrt{\frac{8T\sigma^2}{\pi}} e^{\frac{-x^2}{2T\sigma^2}} \tag{11}$$

Corollary 9 *The expected value of the maximum of μ_{iT}, $1 \leq i \leq n$, is $\theta(\sqrt{T})$.*

Note that in the corollary above n and σ are constant. Now, similar to Lemma 6, we bound $E[\Delta_T]$. The proof is given in appendix A.2.

Lemma 10 *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 11 *$\mathcal{M}_\epsilon(\mathcal{B})$ is ex-post individually rational. Also, for $0 \leq \epsilon \leq \frac{1}{3}$, $\mathcal{M}_\epsilon(\mathcal{B})$ is asymptotically incentive compatible, ex-ante efficient, and has a revenue asymptotically equivalent to the revenue of the efficient second price auction.*

Proof : We first prove ex-post individual rationality. Note that the agents only pay for the items they receive during exploitation. Assume agent i is the person who has received the item at time T and it was during exploitation. i.e., $y_{iT} = 1$. By the payment rule of the mechanism we have:

$$\sum_{t=1}^{T} p_{it} = \sum_{t=1}^{T-1} y_{it} \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_{it}(T)\} \leq \sum_{t=1}^{T-1} y_{it} \widehat{\gamma}_{it}(t)$$

Because $y_{iT} = 1$ we have $\widehat{\gamma}_{it}(t) \leq \widehat{\mu}_{it}(t)$. Therefore we get:

$$\sum_{t=1}^{T} p_{it} \leq \sum_{t=1}^{T-1} y_{it} \widehat{\mu}_{it}(t) = \sum_{t=1}^{T-1} y_{it} r_{il_{i,t-1}} \leq \sum_{t=1}^{T} x_{it} r_{it}.$$

We complete the proof by showing that conditions $(C1)$, $(C2)$, and $(C3)$ hold. By (8), the expected utility of each agent at time t from random exploration is

$$\eta_\epsilon(t)\mu_{it} = \theta(t^{-\epsilon} \ln^{1+\epsilon} t \sqrt{t\sigma^2}) = \theta(t^{\frac{1}{2}-\epsilon} \ln^{1+\epsilon} t). \tag{12}$$

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

Also, by Lemma 10 and Corollary 9:

$$E[\max_{t \leq T}\{\mu_{iT}\} + \sum_{t=1}^{T-1} \Delta_t] = O(T^{1+\frac{\epsilon}{2}}) \tag{13}$$

15

For $\epsilon \leq \frac{1}{3}$, we have $\frac{3}{2} - \epsilon \geq 1 + \frac{\epsilon}{2}$. Therefore, by equations (12) and (13), for $\epsilon \leq \frac{1}{3}$, condition $(C1)$ is met.

By Corollary 9, the expected value of $\max_i\{\mu_{iT}\}$ and γ_T are of $\theta(\sqrt{T})$. Therefore up to time T, both the expected welfare and the expected revenue of the efficient second price mechanism are of $\theta(T^{\frac{3}{2}})$. For any $0 < \epsilon < 1$, we have:

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

Therefore, conditions $(C2)$ and $(C3)$ are satisfied, and $\mathcal{M}_\epsilon(\mathcal{B})$ is asymptotically ex-ante efficient and it has a revenue asymptotically equivalent to the efficient second price auction. $\square$

6   Discussion and Open Problems

In this section we discuss a few extensions of our mechanism.

Multiple Slots: We can use the same idea as Gonen and Pavlov [13], to modify $\mathcal{M}$ when there are multiple slots for showing an ad. Like [13], we need to assume that there exists a set of conditional distributions which determine the probability that the ad in slot j_1 is clicked on conditioned on the event that there was a click on the ad in slot j_2 During exploitation, $\mathcal{M}$ 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 agents realize their value for receiving the item with some delay. 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. Since the computation of the payments in our mechanism is cumulative and adaptive, it is reasonable to believe that our mechanism will maintain its properties as long as the delay in reporting the utilities is bounded. We leave this as an open problem.

Creating Multiple Identities: During exploration, our mechanism gives the item for free to one of the agents chosen uniformly at random. Therefore, it is easy to see that an agent can benefit from participating in the mechanism with multiple identities. This may not be cheap or easy for all advertisers. After all, the traffic should be eventually routed to a legitimate business. Still, a possible solution is increasing the cost of creating new identities by charging advertisers a fixed premium for entering the system (see [18] for a discussion.) Finding a more natural solution remains as an interesting open problem.

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] E. Elkind. Designing And Learning Optimal Finite Support Auctions. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2007.

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

[13] 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.

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

[15] 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.

[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.

A   Missing Proofs of Section 5

A.1   Proof of Lemma 6

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:

$$\Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i]$$

$$= \Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i \wedge n_{it} \geq \frac{E[n_{it}]}{2}]$$

$$+ \Pr[|\mu_i - \widehat{\mu}_{it}(t)| \geq \frac{1}{\sqrt{t^{1-\epsilon}}}\mu_i \wedge n_{it} < \frac{E[n_{it}]}{2}]$$

$$\leq 2e^{\frac{-\frac{1}{t^{1-\epsilon}}t^{1-\epsilon}\ln^{1+\epsilon}t\ \mu_i}{2d}} + e^{\frac{-t^{1-\epsilon}\ln^{1+\epsilon}t}{8d}}$$

$$= o(\frac{1}{t^c}), \forall c > 0$$

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

A.2 Proof of Lemma 10

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:

$$\Pr[T - l_{i,T-1} > t] < (1 - T^{-\epsilon} \ln^{2+2\epsilon} T)^t \leq e^{\frac{-t \ln^{2+2\epsilon} T}{T^\epsilon}}. \tag{14}$$

Let $t = \frac{1}{\ln^{1+\epsilon} T} T^\epsilon$. By equation (10) and (14),

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

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 9, 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 $E[\Delta_T] \leq E[\max_i\{X_{it}\}]$. $\square$