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

Hamid Nazerzadeh* Amin Saberi† Rakesh Vohra‡

September 24, 2012

Abstract

We examine the problem of allocating an item repeatedly over time amongst a set of agents. The value that each agent derives from consumption of the item may vary over time. Furthermore, it is private information to the 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 Pay-Per-Action or Pay-Per-Acquisition (PPA) charging scheme in online advertising. In this scheme, instead of paying per click, advertisers pay only when a user takes a specific action (e.g. purchase an item or fills out a form) on their websites.

1 Introduction

We study a mechanism design problem for the following setting: a number of agents competing for identical items sold repeatedly at times $t = 1, 2, \cdots$. At each time t, the mechanism allocates the item to one of the agents. Agents *discover* their values for the good only if it is allocated to them. If agent i receives the good at time t, she realizes her value, u_{it}, (denominated in money) for it and reports (not necessarily truthfully) the realized value to the mechanism. Then, the mechanism determines how much the agent has to pay for receiving the item. We allow the value of an agent to change over time. An example of such environments, is that of online advertising where advertisers are uncertain about the profit each advertisement opportunity generates. For these environments, we are interested in auction mechanisms which have the following four properties:

1. Agents profit from participating in the mechanism.
2. Agents have an incentive to truthfully report their realized values.
3. The social welfare (and revenue) is, in an appropriate sense, close to a second-price auction.
4. The performance of the mechanism does not depend on an a-priori knowledge of the distribution of u_{it}'s. This is motivated by the Wilson doctrine that criticizes an overdependence

*Marshall School of Business, University of Southern California, Los Angeles, CA, hamidnz@marshall.usc.edu
†Management Science and Engineering Department, Stanford University, Stanford, CA, saberi@stanford.edu
‡Kellogg School of Management, Northwestern University, Evanston, IL, r-vohra@kellogg.northwestern.edu

on common-knowledge assumptions.[1]

The precise manner in which these properties are formalized is described in Section 2. We propose a simple dynamic mechanism that under minimal assumptions satisfies all the above properties *asymptotically* (in time). Key challenges in designing a mechanism in this environment are that (i) the value of an agent for the item may change over time, (ii) the agent may be uncertain about her own value of the item before receiving it, and (iii) the value she obtains after consuming the item will be her *private information*. To address these challenges, we will build our mechanism upon a learning algorithm that estimates the values of the agents based on their reports. Furthermore, our mechanism will incentivize the agents to truthfully report their private information. Roughly speaking, the mechanism allocates the item to the agent with the highest (estimated) reported value. If an agent consistently under-reports her value so that her reported value drops below other agents, then the mechanism will stop allocating the item to her. On the other hand, if the agent over-reports her value and that results in receiving the item more often, she would end up paying more than her actual value.

More specifically, we convert a sampling-based learning algorithm to a mechanism with the above properties by adding an *adaptive second-price scheme*. The mechanism takes two randomly *alternating* actions: exploration and exploitation. During exploration, the item is allocated to a randomly chosen agent for *free*. During exploitation, the mechanism allocates the item to the agent with the highest estimated expected value. After each allocation, the agent who has received the item reports its realized value. Then, the mechanism updates its estimates of the agents' values and determines the payments by setting the price of every item equal to the second-highest *estimated* value. These prices are constantly updated as the mechanism obtains more accurate estimates about the values.

We identify sufficient conditions for a learning algorithm that, in combination with the adaptive second-price scheme, gives a mechanism which asymptotically satisfies the above properties. These conditions can be interpreted as bounds on the exploration rate (the probability of doing exploration at each time t).

Our bounds on the exploration rate require the learning algorithm to obtain reliable estimates of the highest and second-highest valuations with a reasonable exploration rate. We note that the guarantees we obtain on the incentive properties of the mechanism, its social welfare, and revenue are all asymptotic. In particular, they imply that the difference between utilities an agent obtains from a truthful strategy and a deviation from it vanishes as the number of allocated items increases.

A sufficient conditions for the learning algorithm are not very restrictive: we present two concrete examples to show how they can be applied to convert simple learning algorithms to desired mechanisms. In the first example, the value of each agent i for the item at each time t is generated by a Bernoulli random variable with (a-priori unknown) mean μ_i, i.e., $u_{it} = 1$ with probability μ_i and 0 with the remaining probability. We construct our mechanism using a variation of the well-known ϵ-greedy algorithm, see Sutton and Barto (1998). In this setting, the algorithmic question of maximizing the welfare – when agents are *not* strategic – has been extensively studied in

[1] Wilson (1987): "Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent that it assumes other features to be common knowledge, such as one agent's probability assessment about another's preferences or information. I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality."

the context of (non-Bayesian) multi-armed bandits, e.g., see Berry and Fristedt (1985), Auer et al. (2002a), Cesa-Bianchi and Lugosi (2006). In another example, we present a mechanism for a setting where the values of the agents evolve like reflective Brownian motions.

For the problem of selling items to a sequence of short-lived agents, exploration-exploitation based algorithms have been shown to obtain asymptotically optimal revenue (Besbes and Zeevi (2009, 2010)); see also Harrison et al. (2010), Broder and Rusmevichientong (2010). A conceptual contribution of our work is showing that incentive compatibility (and asymptotically optimal welfare and revenue) can be obtained in environments with long-leaved agents, if the exploration rate of the learning algorithm is *independent* of the reports of the agents; this independence is even necessary for more restrictive solution concepts, see Section 1.2.

Organization The rest of the paper is organized as follows: In Section 1.1, we motivate our work in the context of online advertising. In Section 1.2, we briefly review the related literature on dynamic mechanism design. The model and definitions are discussed in Section 2 followed by the description of the mechanism in Section 3. In Section 4, we present our main results and apply them to two examples. We analyze our mechanism in Section 5. In Section 6 we describe a few extensions of our mechanisms. We concluded by discussing future directions of this work.

1.1 Applications to Online Advertising

Online advertising is the main source of revenue for a variety of companies and individuals offering services or content on the web. Essentially, it is a mechanism for "publishers" to sell parts of the space on their webpages to advertisers. A typical advertiser uses this space to attract users to her website with the hope that the user will take an *action* during his visit that generates a value; for instance the user may purchase a product or service or subscribe to a mailing list.

Currently, the dominant model in the online advertising industry is Pay-Per-Click (PPC). In this model, advertisers specify their bids for placing their ads but they only have to pay the publisher if the user clicks on them. A drawback of the PPC 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 for advertisers. This is because the probability that a user takes a profit-generating action after clicking on the ad depends on many factors including the demographics of the user, his past behavior, and content of the webpage.

The problem studied in this paper is a model of a recently popular payment scheme known as the Pay-Per-Action or Pay-Per-Acquisition (PPA) model.[2] Instead of paying per click, the advertiser pays only when a user takes a specific action e.g., purchase an item or a subscription (for a more detailed description, see Google (2007)). PPA allows the advertisers to report their payoff *after* observing the user's action, therefore it eliminates the uncertainty of advertisers and reduces their exposure and computational burden.

From a mechanism design perspective, the fundamental difference between PPC and PPA charging models is that a click on an ad can be observed by both advertiser and publisher. However,

[2] For instance, Marissa Mayer, Googles Vice President of Search Product and User Experience, deemed developing the PPA system as the "Holy Grail" of online advertising, see Spencer (2007).

the action of the user is hidden from the publisher and is observable only by the advertiser. [3] This is the difference that motivates the present paper. The model studied in our paper corresponds to this application as follows: the items being allocated are space on the publisher's page. Advertisers compete for the items, but they are uncertain about their values. In this context, agents are strategic. The second-price auction is a benchmark for measuring the social welfare and revenue of the existing Pay-Per-Click auctions, cf., Edelman et al. (2007). Finally, because of the size and scale of this application, the asymptotic analysis employed in this paper seems quite appropriate. We discuss this in more details in Section 2.

It is worth noting the major difference between PPC and PPA in terms of vulnerability to click fraud. Click fraud refers to clicks generated 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. Click fraud is considered by many experts to be one of the biggest challenges facing the online advertising industry c.f. Grow et al. (2006), Crawford (2004), Mitchell (2005), Immorlica et al. (2005). PPA schemes are less vulnerable to click-fraud 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 a profit.[4]

1.2 Related Work

In this section, we describe a selection of related work on dynamic mechanism design and refer the reader to Bergemann and Said (2011) for a more extensive survey.

We divide the literature into two categories based on the properties of the private information. In the first category, we have a dynamic population of agents but the private information of each agent remains fixed over time. In the second, there is a fixed population but the private information of the agents evolve. Our work belongs to the second category.

Within each category, we further subdivide each section to Bayesian and prior-free settings. In the Bayesian setting, the distribution of private information of the agents and also the underlying dynamics of the environment are common knowledge. This is obviously a very strong assumption. The more recent work on prior-free mechanism design tries to eliminate or minimize these assumption – see Parkes (2007) for a survey.

1.2.1 Fixed Private information

In this category we consider mechanisms for selling a sequence of items over time where agents (buyers) arrive and depart dynamically. For this setting, Parkes and Singh (2003) proposed a dynamic implementation of the VCG mechanism – based on Markov decision processes – that

[3] To address this issue, an approach taken in the industry (e.g., Advertising.com, Admedia.com, Snap.com) is to require the advertisers to install a software that will monitor actions that take place on their web site. For an example of conversion tracking technologies see Google Analytics (http://www.google.com/analytics/.) However, even moderately sophisticated advertisers can find a way to manipulate the software if they find it sufficiently profitable; see Agarwal et al. (2009). See also Dellarocas (2012) for further discussion on the interactions of the advertiser and the publishers under PPA.

[4] Of course in this case, PPA makes generating a fraudulent action a more costly enterprize, but not impossible (one could use a stolen credit card number for example.). We refer the reader to Mahdian and Tomak (2007) for a more detailed discussion on the advantages of PPA over PPC.

maximizes social welfare. For revenue maximization, Vulcano et al. (2002) give a revenue-optimal auction where agents are impatient, i.e., they leave if they do not receive the item upon the arrival, see also Gallien (2006), Gershkov and Moldovanu (2009). For the setting in which agents are patient and their arrival and departure time is their private information, Pai and Vohra (2009) proposed a revenue-optimal mechanism. Recently, Said (2012) and Skrzypacz and Board (2010) show that under certain assumptions, such mechanisms can be implemented via simple auctions or posted-prices. For a similar setting, but without the assumption of having Bayesian priors, (Hajiaghayi et al. 2004, 2005) gave incentive compatible approximately-optimal mechanisms using techniques from online competitive analysis.

Recently, and after the publication of our preliminary results in Nazerzadeh et al. (2008), there have been several results at the intersection of mechanism design and learning motivated by the application of online advertising. Babaioff et al. (2009) consider the mechanism design problem for repeatedly allocating an item (ad space) over a horizon of length T to a set of agents (advertisers): each agent i has a value-per-click v_i which is private information. The value of the agent for the advertisement space is defined as v_i times the probability of the click. The click event corresponds to a Bernoulli random variable with an unknown parameter. The authors propose an incentive compatible mechanism and show that this mechanism achieves the optimal welfare (i.e., minimum regret) among all ex-post truthful (strategy-proof) mechanisms. This is an strong notion of incentive compatibility which requires the truthful strategy to be ex-post dominant. The authors give a characterization of such mechanisms which implies that a strict separation between the exploration and exploitation phases of the algorithm, i.e., the report of agent during an exploitation should not be used by the learning algorithms. More recently Babaioff et al. (2010), improved this result by presenting an ex-ante truthful mechanism which is based on a novel randomized pricing scheme. This mechanism in expectation achieves the same social welfare as the optimal learning algorithm when the agent are non-strategic.

Devanur and Kakade (2009) study a similar setting but they are concerned with maximizing the revenue. Their benchmark for the revenue is the same as ours but they derive a similar characterization to those of Babaioff et al. (2009) for maximizing welfare.

Finally, we remark that none of the above mechanisms would be incentive compatible in the setting considered in this paper. This is because, unlike the above results, the private information in our model is not single dimensional and it evolves over time.

1.2.2 Dynamic Private Information

In this category, we consider models in which agents have repeated interactions with the mechanism and they receive new private information during these interactions. Designing incentive compatible mechanisms is typically much more challenging in such settings.

For the problem of maximizing social welfare, Bergemann and Välimäki (2010) and Athey and Segal (2007) propose elegant dynamic implementations of the efficient mechanisms such as VCG and AGV mechanism (d'Aspremont and Grard-Varet (1979)).

For maximizing revenue, the main idea is to extend the classic revenue-optimal auction of Myerson (1981) to the dynamic settings. Due to the repeated interaction of the agents with the mechanisms, the optimal mechanism usually consists of sequence of screenings; see Courty and Li (2000), Akan et al. (2008). Also, in such dynamic settings, Ëso and Szentes (2007), Pavan et al. (2009), Kakade et al. (2010) showed that a mechanism can obtain more revenue by charging the agents in advance, for the utility they may enjoy in the future. In the context of online advertising,

Kakade et al. (2010) propose a revenue-optimal dynamic mechanism.

In all of the above works, satisfying the incentive compatibility constraints is heavily dependent on the common (Bayesian) prior assumptions. This is a standard but very strong assumption. Even though our work falls into the category of dynamic private information, it differs from the above results because the distributional assumptions are minimal. On the flip side, our mechanisms satisfy the incentive constraint only asymptotically. Namely, we show that the benefit an agent may obtain by deviating from the truthful strategy would vanish over time. The intuition is that the agents will not deviate from the truthful strategy if the benefit of deviation is negligible and finding a profitable deviation is costly. Even in some static environments, as argued by Schummer (2004), if the common Bayesian prior assumptions do not hold, approximate incentive compatibility would circumvent some of the impossibility results. Approximate incentive compatibility implies that the benefit from deviation is bounded; see Lipton et al. (2003), Feder et al. (2007), Daskalakis et al. (2009) on the notion of approximate Nash equilibrium. Asymptotic incentive compatibility has been studied in other contexts as well; for instance, Roberts and Postlewaite (1976), Gul and Postlewaite (1992), and Kojima and Manea (2010) showed that the agents would behave truthfully when the size of the market grows.

2 Definitions and Notation

We consider a setting with the following properties: n agents compete for items allocated repeatedly at periods $t = 1, 2, \cdots$ — one item is sold at each period. The (nonnegative) value of agent i for the item at time t is denoted by u_{it}. Values are denominated in a common monetary scale and they may evolve over time according to a stochastic process. For every pair of agents i and j, $i \neq j$, the evolution of u_{it} and u_{jt} are independent. u_{it} evolves independently of whether or not agent i receives the item at time t. 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 *discovers* u_{it}, her value for the item. Then, agent i, reports r_{it} as her value to the mechanism. The mechanism then determines the payment, denoted by p_{it}. We do not require an agent to know her value for the item before acquiring it. If an agent is not allocated the item, she may not learn any new information about her value. She may also misreport her value to the mechanism after the allocation.

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

We design a mechanism that satisfies the desired properties discussed in the introduction in an *asymptotic* sense. In the context of online advertising, the motivating application for our work, search engines run thousand and sometimes millions of auctions every day to sell the advertisement space for the same keyword. Hence, we expect that a mechanism that asymptotically satisfies the desired properties would perform well in practice.

Individual Rationality: $\mathcal{M}$ is *asymptotically ex-ante individually rational* if the long-term total

utility of agent i, $1 \leq i \leq n$, is non-negative:

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

Incentive Compatibility: We say that a mechanism is asymptotically incentive compatible if truthfulness defines an asymptotic Nash equilibrium, i.e., truthfulness is an asymptotic best response to the truthful strategy of other agents. Formally, consider agent i and suppose all agents except i are truthful. Let $U_i(T)$ be the expected total utility of agent i, if agent i is truthful between time 1 and T.

$$U_i(T) = E\left[\sum_{t=1}^{T} x_{it} u_{it} - p_{it}\right]$$

Also, let $\widetilde{U}_i(T)$ be the maximum of expected profit of agent i under any other strategy when all other agents are truthful. *Asymptotic incentive compatibility* requires that

$$\lim_{T\to\infty} \frac{\widetilde{U}_i(T)}{U_i(T)} = 1$$

We use this notion of incentive compatibility for the following reasons. Since neither the mechanism nor the agents know the distribution of the values, the notation of Bayesian Nash Equilibrium is not applicable. A stronger notion of incentive compatibility, than the one considered here, is the dominant strategy implementation where truthfulness is a utility maximizing strategy, with respect to all the realizations of the agent's own value, the values of other agents, and moreover, the *off equilibrium* behaviors of the other agents in a dynamic setting. This notion appears to be restrictive in dynamic environments and can be implemented only in settings where valuations of the agents evolve *independent* of the allocations of the mechanism, see Kakade et al. (2010), Pavan et al. (2009). An intuition behind the asymptotic (or approximate) incentive compatibility is that calculating a profitable deviation can be difficult and if the increased profit from that strategy is negligible, then the agent does not deviate. Note that in our environment, to calculate the utility maximizing strategy, an agent needs to form a belief about the values and behavior of other agents, and then solve a complicated problem (e.g., a multi-dimensional dynamic program in the case that the belief over the evolution of the valuations corresponds to Markov processes).

Efficiency and Revenue: We measure the efficiency and revenue of $\mathcal{M}$ by comparing it to the second-price mechanism that allocates the item to an agent in $\text{argmax}_i\{\mu_{it}\}$ and charges her the second-highest μ_{it}. The second-price mechanism is ex-ante efficient because, at each time t, it allocates the item to an agent with the highest expected value. The total social welfare obtained by this mechanism up to time T is $E\left[\sum_{t=1}^{T} \max_i\{\mu_{it}\}\right]$. 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 $E\left[\sum_{t=1}^{T} \gamma_t\right]$.

In the setting studied by this paper, where the distributional assumptions are minimal, the best revenue one can hope for is the second-highest price, c.f. Goldberg et al. (2006). In

particular, in the context of online advertising, most publishers try to maximize the efficiency and use the second-highest price as the benchmark for the revenue of their cost-per-click mechanism cf. Edelman et al. (2007).

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.

$$\begin{aligned} W(T) &= E\left[\sum_{t=1}^{T}\sum_{i=1}^{n} x_{it}\mu_{it}\right] \\ R(T) &= E\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{it}\right] \end{aligned}$$

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

$$\lim_{T\to\infty} \frac{E\left[\sum_{t=1}^{T} \max_i\{\mu_{it}\}\right]}{W(T)} = 1.$$

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

$$\lim_{T\to\infty} \frac{E\left[\sum_{t=1}^{T} \gamma_t\right]}{R(T)} = 1.$$

3 A Dynamic Second-price Mechanism

We build our mechanism on top of a learning algorithm, denoted by $\mathcal{L}$, that estimates the expected values of the agents. Let $\widehat{\mu}_{it}\big(u_{i,t_1}, u_{i,t_2}, \cdots, u_{i,t_\kappa}\big) = E_{\mathcal{L}}\left[u_{it}|u_{i,t_1}, u_{i,t_2}, \cdots, u_{i,t_\kappa}\right]$ be the estimation of the learning algorithm of μ_{it} using the realizations $u_{i,t_1}, u_{i,t_2}, \cdots, u_{i,t_\kappa}$ of the agent's value at periods $t_1, t_2, \cdots, t_\kappa$. When it is clear from the context, we drop the subscript. With abuse of notation, we define $\widehat{\mu}_{it}(T) = \widehat{\mu}_{it}\left(\{u_{i,t'}|x_{i,t'} = 1, 1 \le t' < T\}\right)$ to be the estimation of the learning algorithm of μ_{it} conditioned on the realizations of her value when the item was allocated to her, up to but not including, time T. Note that T can be bigger than t. In other words, the learning algorithm may refine its earlier estimates using more recent history. Note that in contrast with the mechanism, an algorithm is defined without concerns of strategic behaviors. The above definitions are with respect to actual valuations of agents not their reports.

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 7). In Sections 4.1.1 and 4.1.2 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[5], 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 value. Then, the agent reports her value to the mechanism and the mechanism determines the payment.

[5] We assume that the exploration rate, η, changes only as the function of time, and not according to the realized values of the agents.

μ_{it}: The expected value of agent i for the item at time t.

$\widehat{\mu}_{it}(T)$: The *estimation* of μ_{it} with information obtained up to time T.

γ_t: The second-highest μ_{it}.

$\widehat{\gamma}_t(T)$: The estimated second-highest price, i.e., $\max_{\{i:x_{it}\neq 1\}} \widehat{\mu}_{it}(T)$.

x_{it}: The indicator variable of allocation; $x_{it} = 1$ iff item is allocated to agent i at time t.

y_{it}: The indicator variable of allocation during exploitation; $y_{it} = 1$ iff $x_{it} = 1$ and t belongs to the exploitation phase.

r_{it}: The report of agent i for her value.

p_{it}: The payment of agent i at time t.

Figure 1: Summary of the notations

We first formalize our assumptions about the learning algorithm and then we discuss the payment scheme. The mechanism is given in Figure 2.

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; y_{it} is defined to be equal to 0 during exploration. The payment of agent i at time t, denoted p_{it}, is equal to:

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

Therefore, we have:

$$\sum_{k=1}^{t} p_{ik} = \sum_{k=1}^{t} y_{ik} \min\{\widehat{\gamma}_k(k), \widehat{\gamma}_k(t)\}. \tag{2}$$

Let us describe the properties of the payment scheme in more detail. First of all, an agent only pays for the items that are allocated to her during exploitation. Second, taking minimum with $\widehat{\gamma}_t(t)$ ensures individual rationality of the mechanism. Because of estimation errors, the mechanism may occasionally allocate the item to an agent with a low expected value. Therefore, by allowing the mechanism to charge $\widehat{\gamma}_t(T)$ after the algorithm corrects its estimates, we can satisfy the individual rationality constraints. See Lemma 11 for a more detailed discussion.

Define $price_{it}(T)$ to be the price of the item allocated at time t computed using the history of reports up to time T. In other words, $price_{it}(T) = \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\}$ where the term $\widehat{\gamma}_t(T)$ corresponds to the up-to-date second-highest estimated price and $\widehat{\gamma}_t(t)$ corresponds to the second-highest estimated price at time t. Since prices change dynamically, the payment of an agent can

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

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

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

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

Figure 2: Mechanism $\mathcal{M}$

be negative sometimes, i.e., the mechanism may pay the agent if she was overcharged earlier. At the same time, the payment of an agent might be positive even if $r_{it} = 0$. This is why the payment scheme of the mechanism is pay-per-action in aggregate where the agent pays for aggregate prices of the items received over time. It is worth noting that the payments and prices in the existing pay-per-click mechanisms are also aggregate over short time scales c.f. Athey and Nekipelov (2010). Later on, in Section 6.1, we show that in certain models it is possible to obtain a stronger notion of ex-post individual rationality.

4 Main Result

In this section, we derive sufficient conditions for a learning algorithm, that when used with a second-price scheme discussed in the last section, yields a mechanism satisfying the desired properties we discussed. Call a learning algorithm *proper*, if the accuracy of its estimation is monotone in the number of samples, that is if for any time $T \geq t$, and subsets $S_1 \subset S_2 \subset \{1, 2, \cdots, T-1\}$, we have

$$E\left[|\widehat{\mu}_{it}\left(\{u_{i,t'}\}_{t' \in S_1}\right) - \mu_{it}|\right] \geq E\left[|\widehat{\mu}_{it}\left(\{u_{i,t'}\}_{t' \in S_2}\right) - \mu_{it}|\right] \tag{3}$$

One more definition is in order before stating our main result.

Definition 2 *Define Δ_t to be the maximum over all agents i, the difference between μ_{it} and its estimate using only reports taken during exploration, i.e.,*

$$\Delta_t = \max_i \Big\{ |\widehat{\mu}_{it}\left(\{u_{i,t'} | x_{i,t'} = 1, y_{i,t'} = 0, 1 \leq t' < t\}\right) - \mu_{it}| \Big\}$$

Theorem 1 *Consider proper learning algorithm $\mathcal{L}$. If the following holds for every agent i and time $T > 0$:*

$$(C1) \qquad E_{\mathcal{L}}\left[\sum_{t=1}^{T} \Delta_t\right] = o\left(E_{\mathcal{L}}\left[\sum_{t=1}^{T} \eta(t)\mu_{it}\right]\right)$$

then mechanism $\mathcal{M}$, as defined in Section 3, is asymptotically individually rational and asymptotically incentive compatible. Also, if in addition to $(C1)$, the following condition holds:

$$(C2) \qquad E_{\mathcal{L}}\left[\sum_{t=1}^{T} \eta(t) \max_i\{\mu_{it}\}\right] = o\left(E_{\mathcal{L}}\left[\sum_{t=1}^{T} \gamma_t\right]\right)$$

then the welfare and revenue of the mechanism are asymptotically equivalent to those of the mechanism that at every time allocates the item to the agent with the highest expected value and charges the winning agent the second-highest expected value.

The proof of the above theorem is given in Section 5. We start by discussing conditions $(C1)$ and $(C2)$. When it is clear from the context, we drop the subscript.

Condition $(C1)$ relates the estimation error of the learning algorithm Δ_t to the exploration rate $\eta(t)$. Interestingly, the exploration rate is related to incentive compatibility and individual rationality: the agents receive the item for free when they are explored, so they get a benefit from participating in the mechanism which, up to time T, sums up to $E\left[\sum_{t=1}^{T} \frac{1}{n}\eta(t)\mu_{it}\right]$. On the other hand, during an exploitation round, an agent may get overcharged by as much as Δ_t because the learning algorithm has overestimated her valuation or the valuation of the second-highest agent. Theorem 1 states that if the total gain from the former effect is asymptotically bigger than the total loss from the latter, then we can achieve individual rationality.

For incentive compatibility, we show that the maximum utility an agent would obtain by deviating from the truthful strategy is bounded by the estimation error of the learning algorithm, namely $O\left(E\left[\sum_{t=1}^{T} \Delta_t\right]\right)$. So when this quantity is asymptotically smaller than the utility obtained by the agent we ensure incentive compatibility as well.

It is also easy to see the intuition behind Condition (C2). During exploration phases, the mechanism allocates the item to a randomly chosen agent for free, incurring a loss of revenue and efficiency. The left hand side term upper bounds the total loss and Condition (C2) requires that it is asymptotically small compared to the maximum revenue of a second-price mechanism.

Note that Conditions $(C1)$ and $(C2)$ suggest how to choose the exploration rate as they give upper and lower bounds. In condition $(C1)$, the right-hand-side is decreasing and the left-hand-side is increasing in the exploration rate. So, one can increase the exploration rate of the algorithm to satisfy this condition. On the other hand, the left-hand-side of Condition $(C2)$ is increasing in the exploration rate, so it gives an upper bound.

4.1 Examples

As two examples of the applications of the theorem, we study two models for the values of the agents. In the first model, the values of the agents are independent and identically-distributed. In the second, the value of each agent evolves independently like a reflected Browning motion. In both of these examples, we give simple sampling-based algorithms for learning the values of the agents and show how we can adjust the exploration rates to satisfy the conditions of Theorem 1.

4.1.1 Independent and Identically-Distributed Values

Assume that for each i, u_{it}'s are independent and identically distributed random variables. For simplicity, we define $\mu_i = E[u_{it}], t > 0, 0 < \mu_i \leq 1$.

In this environment, the learning algorithm we use is an ε-greedy algorithm for the multi-armed bandit problem cf. Sutton and Barto (1998). For $\epsilon \in (0,1)$, define:

$$\begin{aligned} n_{it} &= \sum_{k=1}^{t-1} x_{it} \\ \eta_\epsilon(t) &= \min\{1, nt^{-\epsilon}\ln^{1+\epsilon} t\} \\ \widehat{\mu}_{it}(T) &= \begin{cases} (\sum_{k=1}^{T} x_{ik} r_{ik})/n_{iT}, & n_{iT} > 0 \\ 0, & n_{iT} = 0 \end{cases} \end{aligned} \tag{4}$$

Call the mechanism based on this learning algorithm $\mathcal{M}_\epsilon(iid)$. Note that $\widehat{\mu}_{it}(T)$ is the average of the reports of the agent i up to, but not including, time T.

Using the Chernoff bound, it is easy to see that the learning algorithm is proper. We prove the following lemma, see Appendix A.

Lemma 2 *If all agents are truthful, then for $\mathcal{M}_\epsilon(iid)$, we have $E[\Delta_t] = O\left(\frac{1}{\sqrt{t^{1-\epsilon}}}\right)$.*

We remark that randomization is not essential and it suffices that the algorithm explores such that $E[n_{it}] = \theta\left(t^{1-\epsilon}\ln^{1+\epsilon}\right)$.

Lemma 3 *If all agents are truthful, then $\mathcal{M}_\epsilon(iid)$ satisfies Conditions $(C1)$ and $(C2)$ for $\epsilon \leq \frac{1}{3}$.*

Proof : Proof: By the lemma above, $E\left[\sum_{t=1}^{T}\Delta_t\right] = O\left(T^{\frac{1+\epsilon}{2}}\right)$. Also, because $\mu_i \leq 1$, we have $E\left[\sum_{t=1}^{T}\eta_\epsilon(t)\mu_i\right] = \Omega\left(T^{1-\epsilon}\ln^{1+\epsilon} T\right)$. Condition $(C1)$ holds because $\frac{1+\epsilon}{2} \leq 1-\epsilon$ for $\epsilon \leq \frac{1}{3}$.

On the other hand $E\left[\sum_{t=1}^{T}\gamma_t\right] = \theta(T)$. Condition $(C2)$ holds because the exploration rate approaches 0. □

Corollary 4 *By Theorem 1, $\mathcal{M}_\epsilon(iid)$ is asymptotically individually rational and incentive compatible and has efficiency and revenue asymptotically equivalent to those of the efficient second-price auction.*

4.1.2 Brownian Motion

In this section, we consider a setting where for each i, $1 \leq i \leq n$, the evolution of u_{it} is a reflected Brownian motion with mean zero and variance σ_i^2. The reflection barrier is at 0 to avoid negative values. In addition, we assume $\mu_{i0} = 0$ and $\sigma_i^2 \leq \sigma^2$, for some constant σ. In contrast with the previous example, we assume $u_{it} = \mu_{it}$ (but μ_{it} is changing over time).

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.

$$\begin{aligned}\eta_\epsilon(t) &= \min\{1, nt^{-\epsilon}\ln^{2+2\epsilon} t\}\\ \widehat{\mu}_{it}(T) &= \begin{cases} r_{il_{it}} & t<T\\ r_{il_{i,t-1}} & t=T\\ r_{il_{i,T}} & t>T\end{cases}\end{aligned}$$

Call this mechanism $\mathcal{M}_\epsilon(\mathcal{B})$.

We begin by stating some well-known properties of reflected Brownian motions, cf. Borodin and Salminen (2002).

Lemma 5 *Let $[W_t, t\geq 0]$ be a reflected Brownian motion with mean zero and variance σ^2; the reflection barrier is 0. For $T>0$, we have*

$$E[W_t] = \theta\left(\sqrt{t\sigma^2}\right) \tag{5}$$

let $z=$.

$$\Pr[(W_{t+T}-W_t)\in dx] \leq \sqrt{\frac{2}{\pi T\sigma^2}}e^{\frac{-x^2}{2T\sigma^2}} \tag{6}$$

$$\Pr[|W_{t+T}-W_t|\geq x] \leq \sqrt{\frac{8T\sigma^2}{\pi}}\frac{1}{x}e^{\frac{-x^2}{2T\sigma^2}} \tag{7}$$

$$E[|W_{t+T}-W_t|I(|W_{t+T}-W_t|\geq x)] \leq \sqrt{\frac{8T\sigma^2}{\pi}}e^{\frac{-x^2}{2T\sigma^2}} \tag{8}$$

Here $I(.)$ is the indicator function, e.g., $I(|W_{t+T}-W_t|\geq x)=1$ if and only if $|W_{t+T}-W_t|\geq x$.

Corollary 6 *Since $\min_{t_1\in S_1}\{|t-t_1|\}\geq \min_{t_2\in S_2}\{|t-t_2|\}$, the learning algorithm is proper.*

Corollary 7 *The expected value of $\max_{1\leq i\leq n}\{\mu_{iT}\}$ is $\theta\left(\sqrt{T}\right)$.*

Note that in the above corollary n and σ are constant. Now, similar to Lemma 3, we bound $E[\Delta_T]$. The proof is given in Appendix A.

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

Note that $E[\Delta_T]$ is increasing, but it does so at a lower rate compared to $\theta\left(\sqrt{T}\right)$ which, by the above lemma, is the expected value of $\max_{1\leq i\leq n}\{\mu_{iT}\}$.

Lemma 9 *If all agents are truthful, then $\mathcal{M}_\epsilon(\mathcal{B})$ satisfies Conditions $(C1)$ and $(C2)$ for $\epsilon\leq\frac{1}{3}$.*

Proof : Proof: By Eq.(5), the expected value of each agent from exploration and at time t is

$$\eta_\epsilon(t)\mu_{it} = \theta\left(t^{-\epsilon}\ln^{1+\epsilon} t\sqrt{t\sigma^2}\right) = \theta\left(t^{\frac{1}{2}-\epsilon}\ln^{1+\epsilon} t\right). \tag{9}$$

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

Also, by Lemma 8 and Corollary 7, we have

$$E\left[\sum_{t=1}^{T}\Delta_t\right] = O\left(T^{1+\frac{\epsilon}{2}}\right). \tag{10}$$

For $\epsilon \leq \frac{1}{3}$, we have $\frac{3}{2}-\epsilon \geq 1+\frac{\epsilon}{2}$. Thus, by Eq.(9) and (10), for $\epsilon \leq \frac{1}{3}$, Condition $(C1)$ is met.

By Corollary 7, the expected value of $\max_i\{\mu_{iT}\}$ and γ_T are of $\theta\left(\sqrt{T}\right)$. Therefore up to time T, both the expected welfare and the expected revenue of the efficient second-price mechanism are of $\theta\left(T^{\frac{3}{2}}\right)$. For any $0<\epsilon<1$, we have $\theta\left(T^{\frac{3}{2}}\right) = \omega\left(T^{1+\frac{\epsilon}{2}}\right)$, which implies that Condition $(C2)$ is satisfied. □

Corollary 10 *By Theorem 1, $\mathcal{M}_\epsilon(\mathcal{B})$ is asymptotically individually rational and incentive compatible and has efficiency and revenue asymptotically equivalent to those of the efficient second-price auction.*

We believe by adding the formal definition of $\widehat{\mu}$ (see the response to points 7 and 8) it is now clear how we obtain the equation. Unfortunately, the second part of the comment is not clear to us, since neither the valuation model, nor the algorithm are specified. But if we understand correctly, learning algorithms that only use a window of time to estimate valuations would fit in our framework, but they may not satisfy the asymptotic assumptions if the valuations are fluctuation rapidly.

5 Analysis

In this section we prove Theorem 1. In the previous section, we defined Δ_t to be the maximum over all agents i, the difference between μ_{it} and its estimation using only reports taken during exploration (assume all agents are truthful up to time t). Because the accuracy of the learning algorithm is monotone, at time $T \geq t$ we have:

$$E\left[|\widehat{\mu}_{it}(T)-\mu_{it}|\right] \leq E\left[|\widehat{\mu}_{it}(t)-\mu_{it}|\right] \leq E\left[\Delta_t\right] \tag{11}$$

The next lemma relates the individual rationality of the mechanism to the estimation error of the learning algorithm.

Lemma 11 *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:*

$$E\left[\sum_{t=1}^{T}p_{it}\right] - E\left[\sum_{t=1}^{T}y_{it}\mu_{it}\right] \leq E\left[\sum_{t=1}^{T}\Delta_t\right]$$

Proof : Proof: By Eq. (2) we have

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

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

Since y_{it} indicates whether there was an allocation to i at time t during an exploitation phase, $y_{it} = 1$ implies $\widehat{\gamma}_t(t) \leq \widehat{\mu}_{it}(t)$. Plugging that into the inequality above, we get

$$E\left[\sum_{t=1}^{T} p_{it}\right] - E\left[\sum_{t=1}^{T} y_{it}u_{it}\right] \leq E\left[\sum_{t=1}^{T} (\widehat{\mu}_{it}(t) - \mu_{it})^+\right] \leq E\left[\sum_{t=1}^{T} \Delta_t\right]. \tag{12}$$

The last inequality follows from Eq.(11). □

The next lemma shows that if $\sum \Delta_t$ is small, then agents cannot gain much by deviating from the truthful strategy.

Lemma 12 *If all other agents are truthful, agent i cannot increase her expected value up to time T by more than $4E\left[\sum_{t=1}^{T} \Delta_t\right]$ by being non-truthful.*

Proof : Proof Sketch: The complete proof is given in the appendix. Here we give a sketch. First, we bound the expected profit of i for deviating from the truthful strategy. Let $\mathcal{S}$ be the strategy that i deviates to. Fix 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, $D_T = \{t < T | y_{it} = 1, \text{ if the strategy of } i \text{ is } \mathcal{S}\}$. Similarly, let $C_T = \{t < T | y_{it} = 1, \text{ if } i \text{ is truthful}\}$. Using Eq. (11), we show that at time $t \in D_T \cap C_T$, the difference between the prices of the item for agent i under the two strategies is $O(\Delta_t)$. Also, we show that at time $t \in D_T \setminus C_T$, the difference between the expected value and price of the item for the agent i is of $O(\Delta_t)$. In other words, the expected utility an agent would obtain at $t \in D_T \setminus C_T$ is of $O(\Delta_t)$, which completes the proof. □

Using the above lemmas, we can prove that the mechanism is individually rational and incentive compatible.

Lemma 13 *If Condition $(C1)$ holds for the learning algorithm, then the mechanism is asymptotically individually rational and incentive compatible.*

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

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

Hence, for the utility of agent i we have

$$\begin{aligned} U_i(T) &= E\left[\sum_{t=1}^{T} x_{it}u_{it}\right] - E\left[\sum_{t=1}^{T} p_{it}\right] \\ &= \frac{1}{n}E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right] + (E\left[\sum_{t=1}^{T} y_{it}\mu_{it}\right] - E\left[\sum_{t=1}^{T} p_{it}\right]) \\ \text{(by Lemma 11)} \quad &\geq \frac{1}{n}E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right] - E\left[\sum_{t=1}^{T}\Delta_t\right] \\ \text{(by Condition } (C1)) \quad &\geq (\frac{1}{n} - o(1))E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right] \end{aligned} \tag{13}$$

Therefore, the mechanism is asymptotically ex-ante individually rational. Moreover, inequality (13) implies that the utility of the agent i is $\Omega\left(E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right]\right)$. Thus, by Lemma 12, if $(C1)$ holds, then the mechanism is asymptotically incentive compatible. □

Lemma 14 *If Conditions* $(C1)$ *and* $(C2)$ *hold, then the efficiency and revenue of the mechanism are asymptotically asymptotically equivalent to those of the efficient second-price auction.*

Proof : Proof Sketch: (the complete proof is given in the appendix) The mechanism loses welfare and revenue both during exploration and exploitation. In both cases, the lose during exploration is bounded by $E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right]$. We also show that the loss during exploration is bounded by $O\left(E\left[\sum_{t=1}^{T}\Delta_t\right]\right) = o\left(E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right]\right)$ (by Condition $(C1)$). Then, the claim follows from Condition $(C2)$. □

6 Extensions

In this section we consider three modifications of the mechanism.

6.1 Ex-post Individual Rationality

In this section, we use the special structure of the I.I.D and Brownian motion models to prove a stronger result by slightly modifying the payment scheme of the mechanism. A mechanism is *ex-post individually rational* if for any agent i and any sequence of reports up to time $T \geq 1$:

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

This property implies that the total payment of a (truthful) agent would never exceed the value obtained by the agent. We modify the payment scheme of the mechanisms as follows:

$$p_{it} = \sum_{k=1}^{t-1} y_{ik}\min\{\widehat{\gamma}_k(k), \widehat{\gamma}_k(t)\} - \sum_{k=1}^{t-2} p_{ik}. \tag{14}$$

The difference with the original payments scheme is that the agent does *not* pay for the item she receives at time t; she pays only for the item received (during an exploitation) up to, but not including, time t.

Lemma 15 *$\mathcal{M}_\epsilon(iid)$ with the modified payment scheme is ex-post individually rational.*

Proof : Proof: It suffices to prove the claim only for the periods that agent i has received the item during exploitation (i.e., $y_{iT} = 1$), because an agents only pays for these items. Therefore, by Eq. (14), we have:

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

Since y_{it} indicates whether there was an allocation to i at time t during an exploitation phase, $y_{it} = 1$ implies $\widehat{\gamma}_t(t) \leq \widehat{\mu}_{it}(t)$. Thus, we have

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

The last inequality follows from the definition of the learning algorithm Eq. (4). Therefore the mechanism is ex-post individually rational. □

Similarly, we have.

Lemma 16 *$\mathcal{M}_\epsilon(\mathcal{B})$ with the modified payment scheme is ex-post individually rational.*

Proof : Proof: Assume agent i is the person who has received the item at time T and it was during exploitation, i.e., $y_{iT} = 1$. Similar to Lemma 15, we have:

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

which completes the proof. □

We now show that the modified mechanism still satisfies the desired properties. Because the agent does not pay for the last item allocated to her, then by Lemma 12, the maximum expected utility an agent may obtain, up to time T, by deviating from the truthful strategy is bounded by $E\left[\max_{t\leq T} \mu_{it}\right] + O\left(E\left[\sum_{t=1}^{T} \Delta_t\right]\right)$. Hence, to prove incentive compatibility of the mechanism, it suffices to have $E\left[\max_{t\leq T}\{\mu_{it}\}\right] = o\left(E\left[\sum_{t=1}^{T} \eta(t)\mu_{it}\right]\right)$. Also, note that there is no loss of welfare due to the modification of the payment scheme. Regarding the revenue, the loss of revenue is equal to the sum of the outstanding payments (for the last items allocated to the agents). Hence, this loss up to time T is of $O\left(E\left[\max_{t\leq T}\{\mu_{it}\}\right]\right)$. Therefore, the mechanism satisfies all the desired properties if, in addition to Conditions $(C1)$ and $(C2)$, we have $E\left[\max_{t\leq T}\{\mu_{it}\}\right] = o\left(E\left[\sum_{t=1}^{T} \eta(t)\mu_{it}\right]\right)$.

For the I.I.D setting and mechanism $\mathcal{M}_\epsilon(iid)$, $\epsilon < 1$, we have

$$E\left[\max_{t\leq T}\{\mu_{it}\}\right] = \mu_i = \theta(1) = o\left(T^{1-\epsilon} \ln^{1+\epsilon} T\right) = o\left(E\left[\sum_{t=1}^{T} \eta(t)\mu_{it}\right]\right)$$

For the Brownian motion setting and mechanism $\mathcal{M}_{\mathcal{B}}(iid)$, $\epsilon < 1$, we have

$$E\left[\max_{t\leq T}\{\mu_{it}\}\right] = E\left[\mu_{iT}\right] = \theta\left(\sqrt{T}\right) = o\left(T^{\frac{3}{2}-\epsilon}\right) = o\left(E\left[\sum_{t=1}^{T}\eta(t)\mu_{it}\right]\right)$$

6.2 Multiple Slots

In sponsored search, the advertisement space associated with each search query is usually allocated to multiple advertisers. The model studied in this paper can be easily generalized to this setting under some assumptions. Suppose there are k slots for advertisement. The standard assumption is that the probabilities of click are separable, cf. Edelman et al. (2007); namely, there are coefficients $1 = \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$, such that if the probability of a click on an ad on slot 1 is equal to p, then the probability of a click on the slot $\ell > 1$ is equal to $\alpha_\ell p$.

This corresponds to a setting where the mechanism allocates k items at each time t, and the value of the agent for the ℓ^{th} item is equal to $\alpha_l u_{it}$ – the mechanism allocates at most one item to each agent. The extension of our mechanism to such a setting is as follows: during exploration (with probability $\eta(t)$), the items are randomly allocated to the agents. During exploitation, the mechanism allocates the items to the advertisers with the k highest $\widehat{\mu}_{it}(t)$, e.g., the advertiser with the highest $\widehat{\mu}_{it}(t)$ receives the first item. The prices are determined according to Holmstrom's lemma (see Aggarwal et al. (2006)) which is an (incentive-compatible) extension of the second-price mechanism to this setting. Suppose agent i has received item ℓ at time t during exploitation. The price of the agent for that item, computed at time T, is equal to $\min\{\widehat{\gamma}_t^\ell(t), \widehat{\gamma}_t^\ell(T)\}$ where $\widehat{\gamma}_t^\ell(T)$ is defined as follows. To simplify the exposition, without loss of generality, assume that at time T, for agent $j \neq i$, we have $\widehat{\mu}_{1,t}(T) \geq \cdots \geq \widehat{\mu}_{i+1,t}(T) \geq \widehat{\mu}_{i-1,t}(T) \geq \widehat{\mu}_{n,t}(T)$.

$$\widehat{\gamma}_t^\ell(T) = \sum_{j=\ell}^{k} \widehat{\mu}_{jt}(T)(\alpha_j - \alpha_{j+1})$$

where α_{k+1} is defined to be equal to 0. Note that if k is equal to 1, then $\widehat{\gamma}_{it}^1(T) = \widehat{\gamma}_t(T)$.

Theorem 1 can be (easily) extended to the setting with multiple items. The mechanism preserves its desired properties essentially because the accuracy of estimation would only increase due to an increase in the number of allocations. The key observation is similar to Lemma 12, because $\widehat{\gamma}_t^\ell(T)$ is a linear function of $\widehat{\mu}_{jt}(T)$'s, its estimation error is bounded by $O\left(\Delta_t\right)$.

6.3 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 values and the estimates that the learning algorithm computes from them. This may be advantageous when the bidders themselves are unaware of their expected values. However, sometimes the agents have better estimates of their values 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 $\max_{j \neq i}\{b_{jt}\}$. In this setting, we 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 $E\left[|b_{it} - \mu_{it}|\right] \leq E\left[|\widehat{\mu}_{it}(t) - \mu_{it}|\right]$.

Note that the second condition does not require that agents bid their actual value, only that their bids are closer to their actual values than our estimates. With these modifications, our results in the previous sections continue to hold.

7 Discussion and Open Problems

Other Learning algorithms: In this work, we used *sampling-based* learning algorithms for designing incentive-compatible mechanisms. An important direction would be to design mechanisms using other types of learning algorithms that may have a better performance. For example, in the i.i.d. setting discussed in Section 4.1.1, the regret of $\mathcal{M}_\epsilon(iid)$, for $\epsilon = \frac{1}{3}$, is $O\left(T^{\frac{2}{3}} \log^{\frac{4}{3}} T\right)$. However, the optimal achievable regret is $\Omega\left(\sqrt{T}\right)$ (Auer et al. (2002b), Audibert and Bubeck (2009)). Regret is a standard benchmark for evaluating learning algorithms. In our setting, it measures the difference between the welfare obtained by the algorithm and the optimal algorithm that always allocates the item to an agent with the highest expected value.

An example of such algorithm is that of Auer et al. (2002a) whose regret is very close to optimum (i.e., $O\left(\sqrt{T \log(T)}\right)$. Babaioff et al. (2010) build on that algorithm to design an ex-post incentive compatible learning algorithm when agents have single-dimensional private information (see Section 1.2.1). An interesting open question is that whether this bound can be achieved when agent have *dynamic private information*.

As another example, for the setting where the valuations of the agents evolve according to reflected Brownian motions, with reflecting barriers both from below and above, Slivkins and Upfal (2008) give a learning algorithm. It is not obvious how to design incentive compatible dynamic mechanisms based on these algorithm especially because the exploration rate depends on the reports of the agents.

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, or for the initial impressions. A similar problem emerges even if an advertiser is allowed to define multiple actions. Agarwal et al. (2009) showed that multiple actions provide incentive for the advertiser to skew the bidding. Finding a more natural solution for these problems remains a challenging open problem.

Acknowledgment We would like to acknowledge Peter Glynn for fruitful discussions on the properties of the Brownian motions and also the anonymous referees whose comments and suggestions significantly enhanced the presentation of the paper.

A Omitted Proofs

Proof : Proof of Lemma 2: We prove the lemma by showing that for any agent i,

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

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

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

By the Chernoff-Hoeffding bound:

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

Inequality (11) and the Chernoff-Hoeffding bound imply:

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

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

Proof : Proof of Lemma 8: Define $X_{it} = |\mu_{i,T} - \mu_{i,T-t}|$. We first prove $\Pr\left[X_{it} > T^{\frac{\epsilon}{2}}\right] = o\left(\frac{1}{T^c}\right), \forall c > 0$. 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:

$$\Pr[T - l_{i,T-1} > t] < (1 - T^{-\epsilon}\ln^{2+2\epsilon} T)^t \leq e^{\frac{-t\ln^{2+2\epsilon} T}{T^\epsilon}}. \tag{15}$$

Let $t = \frac{1}{\ln^{1+\epsilon} T} T^{\epsilon}$. By equation (7) and (15),

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

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 $\theta\left(\sqrt{T}\right)$. Therefore, $E\left[\max_i\{X_{it}\}\right] = O\left(T^{\frac{\epsilon}{2}}\right)$. The lemma follows because $E\left[\Delta_T\right] \leq E\left[\max_i\{X_{it}\}\right]$. □

Proof : Proof of Lemma 12: We bound the expected utility of agent i for deviating from the truthful strategy. Let $\mathcal{S}$ be the strategy that i deviates to. Fix 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, is equal to:

$$\begin{aligned}
& E\left[\sum_{t\in D_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] - E\left[\sum_{t\in C_T} \mu_{it} - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\}\right] \\
= \; & E\left[\sum_{t\in D_T\cap C_T} \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] + E\left[\sum_{t\in D_T\setminus C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] \\
\leq \; & E\left[\sum_{t\in D_T\cap C_T} \widehat{\gamma}_t(t) - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] + E\left[\sum_{t\in D_T\setminus C_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] \qquad (16)
\end{aligned}$$

We consider two cases:

1. The first case is when $t \in D_T \cap C_T$. We show that in those cases the "current price" given to agent i under the two scenarios are close. Let $j \neq i$ be the agent with the highest $\widehat{\mu}_{jt}(t)$, i.e., $\widehat{\mu}_{jt}(t) = \max_{j\neq i}\{\widehat{\mu}_{jt}(t)\} = \widehat{\gamma}_t(t)$. Because i is the winner both in D_T and C_T, by the definition, we have $\widehat{\mu}'_{jt}(t) \leq \widehat{\gamma}'_t(t)$ and $\widehat{\mu}'_{jt}(T) \leq \widehat{\gamma}'_t(T)$. Therefore, we get

$$\begin{aligned}
\widehat{\gamma}_t(t) - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} &\leq \widehat{\mu}_{jt}(t) - \min\{\widehat{\mu}'_{jt}(t), \widehat{\mu}'_{jt}(T)\} \\
&\leq \max_{k\neq i}\left\{\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\}\right\} \qquad (17)
\end{aligned}$$

2. The second case is when $t \in D_T \setminus C_T$. We show in those cases agent i cannot increase her utility by much. Let j be the agent who would receive the item at time t when agent i is

truthful. Hence, $\widehat{\mu}_{jt}(t) \geq \widehat{\mu}_{it}(t)$. Therefore, we have

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

Also, $\widehat{\mu}'_{jt}(t) \leq \max_{j \neq i}\{\widehat{\mu}'_{jt}\} = \gamma'_t(t)$. Similarly, $\widehat{\mu}'_{jt}(T) \leq \gamma'_t(T)$. Plugging into the above inequality we get:

$$\begin{aligned}\mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\} &\leq (\mu_{it} - \widehat{\mu}_{it}(t)) + \widehat{\mu}_{jt}(t) - \min\{\widehat{\mu}'_{jt}(t), \widehat{\mu}'_{jt}(T)\} \\ &\leq (\mu_{it} - \widehat{\mu}_{it}(t)) + \max_{k \neq i}\left\{\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\}\right\} \quad (18)\end{aligned}$$

Plugging Eq. (17) and (18) into Eq. (16), we get

$$\begin{aligned}& E\left[\sum_{t \in D_T} \mu_{it} - \min\{\widehat{\gamma}'_t(t), \widehat{\gamma}'_t(T)\}\right] - E\left[\sum_{t \in C_T} \mu_{it} - \min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\}\right] \\ \leq\ & E\left[\sum_{t \in D_T \cap C_T} \max_{k \neq i}\left\{\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\}\right\}\right] \\ & + E\left[\sum_{t \in D_T \setminus C_T} (\mu_{it} - \widehat{\mu}_{it}(t)) + \max_{k \neq i}\left\{\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\}\right\}\right] \\ =\ & E\left[\sum_{t \in D_T} \max_{k \neq i}\left\{\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\}\right\}\right] \\ & + E\left[\sum_{t \in D_T \setminus C_T} (\mu_{it} - \widehat{\mu}_{it}(t))\right] \quad (19)\end{aligned}$$

We start with the first term in the last expression.

$$\begin{aligned}\widehat{\mu}_{kt}(t) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T) &= \mu_{kt} - (\mu_{kt} - \widehat{\mu}_{kt}(t)) - \min\{\widehat{\mu}'_{kt}(t), \widehat{\mu}'_{kt}(T)\} \\ &= \mu_{kt} - (\mu_{kt} - \widehat{\mu}_{kt}(t)) + \max\{-\widehat{\mu}'_{kt}(t), -\widehat{\mu}'_{kt}(T)\} \\ &= \max\{\mu_{kt} - \widehat{\mu}'_{kt}(t), \mu_{kt} - \widehat{\mu}'_{kt}(T)\} - (\mu_{kt} - \widehat{\mu}_{kt}(t)) \\ &\leq |\mu_{kt} - \widehat{\mu}'_{kt}(t)| + |\mu_{kt} - \widehat{\mu}'_{kt}(T)| + |\mu_{kt} - \widehat{\mu}_{kt}(t)|\end{aligned}$$

Hence,

$$\begin{aligned}
& E\left[\sum_{t\in D_T}\max_{k\neq i}\left\{\widehat{\mu}_{kt}(t)-\min\{\widehat{\mu}'_{kt}(t),\widehat{\mu}'_{kt}(T)\}\right\}\right] \\
\leq\ & E\left[\sum_{t=1}^{T}\max_{k\neq i}\left\{|\mu_{kt}-\widehat{\mu}'_{kt}(t)|+|\mu_{kt}-\widehat{\mu}'_{kt}(T)|+|\mu_{kt}-\widehat{\mu}_{kt}(t)|\right\}\right] \\
\leq\ & E\left[\sum_{t=1}^{T}\max_{k\neq i}\left\{|\mu_{kt}-\widehat{\mu}'_{kt}(t)|\right\}+\max_{k\neq i}\left\{|\mu_{kt}-\widehat{\mu}'_{kt}(T)|\right\}+\max_{k\neq i}\left\{|\mu_{kt}-\widehat{\mu}_{kt}(t)|\right\}\right] \\
\leq\ & 3E\left[\sum_{t=1}^{T}\Delta_t\right] \qquad (20)
\end{aligned}$$

The last inequality follows from Eq. (11) because all agents, except i, are truthful. Similarly, we have

$$E\left[\sum_{t\in D_T\setminus C_T}(\mu_{it}-\widehat{\mu}_{it}(t))\right] \leq E\left[\sum_{t=1}^{T}\Delta_t\right] \qquad (21)$$

Therefore, plugging Eq. (20) and (21) into Eq. (19), we get

$$E\left[\sum_{t\in D_T}\mu_{it}-\min\{\widehat{\gamma}'_t(t),\widehat{\gamma}'_t(T)\}\right]-E\left[\sum_{t\in C_T}\mu_{it}-\min\{\widehat{\gamma}_t(t),\widehat{\gamma}_t(T)\}\right]\leq 4E\left[\sum_{t=1}^{T}\Delta_t\right]$$

which completes the proof. □

Proof : Proof of Lemma 14: We first prove the claim for the efficiency and then for the revenue of the mechanism. Our mechanism may lose on welfare both during exploration and exploitation. During exploration, the item is randomly allocated to one of the agents. The expected loss in this case is at most $E\left[\sum_{t=1}^{T}\eta(t)\max_i\{\mu_{it}\}\right]$.

The mechanism can also make a mistake during exploitation: the error in the estimations 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, i.e., $\mu_{it}>\mu_{jt}$. By the rule of the mechanism we have $\widehat{\mu}_{it}(t)\leq\widehat{\mu}_{jt}(t)$. By subtracting this inequality from $\mu_{it}>\mu_{jt}$ we get:

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

We sum up this inequality over all such time t, and by inequality (11), the expected efficiency loss during exploration is bounded by $2E\left[\sum_{t=1}^{T}\Delta_t\right]$.

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

$$\begin{aligned}
E\left[\sum_{t=1}^{T}\max_i\{\mu_{it}\}\right] - W(T) &\leq E\left[\sum_{t=1}^{T}\eta(t)\max_i\{\mu_{it}\}\right] + 2E\left[\sum_{t=1}^{T}\Delta_t\right] \\
\text{(By Condition } (C1)) \qquad &= O\big(E\left[\sum_{t=1}^{T}\eta(t)\max_i\{\mu_{it}\}\right]\big) \\
\text{(By Condition } (C2)) \qquad &= o\big(E\left[\sum_{t=1}^{T}\gamma_t\right]\big) \\
&= o\big(E\left[\sum_{t=1}^{T}\max_i\{\mu_{it}\}\right]\big)
\end{aligned}$$

which leads to the claim for the efficiency.

We now consider the revenue of our mechanism. Similar to the welfare, our mechanism may lose revenue during exploration – bounded by $E\left[\sum_{t=1}^{T}\eta(t)\max_i\{\mu_{it}\}\right]$ – and also during exploitation due to the estimation errors 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 value, then let j be the agent with the highest expected value. The estimation error of γ_t is equal to $\min\{\widehat{\gamma}_t(t), \widehat{\gamma}_t(T)\}$. Observe that $\gamma_t = \mu_{it} \leq \mu_{jt}$, $\widehat{\mu}_{jt}(t) \leq \widehat{\gamma}_t(t)$, and $\widehat{\mu}_{jt}(T) \leq \widehat{\gamma}_t(T)$. Therefore we have,

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

   Therefore in this case, by inequality (11), 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 value. Similar to the previous case, we have

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

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

For the expected revenue of the mechanism we have:

$$\begin{aligned}
R(T) = E\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{it}\right] &\geq E\left[\sum_{t=1}^{T}(1-\eta(t))\min\{\widehat{\gamma}_t(T), \widehat{\gamma}_t(t)\}\right] \\
&\geq E\left[\sum_{t=1}^{T}(1-\eta(t))(\gamma_t - 2\Delta_t)\right] \\
&\geq E\left[\sum_{t=1}^{T}\gamma_t\right] - E\left[\sum_{t=1}^{T}\eta(t)\gamma_t\right] - E\left[\sum_{t=1}^{T}2\Delta_t\right] \\
\text{(by Condition } (C1)) \qquad &\geq E\left[\sum_{t=1}^{T}\gamma_t\right] - (1+o(1))E\left[\sum_{t=1}^{T}\eta(t)\max_i\{\mu_{it}\}\right] \\
\text{(by Condition } (C2)) \qquad &= (1-o(1))E\left[\sum_{t=1}^{T}\gamma_t\right]
\end{aligned}$$

which completes the proof. □

References

Nikhil Agarwal, Susan Athey, and David Yang. Skewed bidding in pay per action auctions for online advertising. *American Economic Review Papers and Proceedings*, 2009.

Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. Truthful auctions for pricing search keywords. In *ACM Conference on Electronic Commerce*, pages 1–7, 2006.

Mustafa Akan, Baris Ata, and James Dana. Revenue management by sequential screening. *Working Paper*, 2008.

Susan Athey and Denis Nekipelov. Equilibrium and uncertainty in sponsored search advertising. *Working paper*, 2010.

Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Working paper*, 2007.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Learning Theory - COLT*, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *ACM Conference on Electronic Commerce*, 2009.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. In *ACM Conference on Electronic Commerce*, pages 43–52, 2010.

Dirk Bergemann and Maher Said. Dynamic auctions: A survey. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78:771–789, 2010.

Donald Berry and Bert Fristedt. *Bandit problems.* London: Chapman and Hall, 1985.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: risk bounds and near-optimal algorithms. *Operations Research*, 57:1407–1420, 2009.

Omar Besbes and Assaf Zeevi. Blind network revenue management. *Working Paper*, 2010.

Andrei Borodin and Paavo Salminen. *Handbook of Brownian Motion: Facts and Formulae.* Springer, 2002.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Working paper*, 2010.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, NY, 2006.

Pascal Courty and Hao Li. Sequential screening. *Review of Economic Studies*, 67:697–717, 2000.

Krysten Crawford. Google cfo: Fraud a big threat. *CNN/Money*, December 2 2004.

Constantinos Daskalakis, Aranyak Mehta, and Christos H. Papadimitriou. A note on approximate nash equilibria. *Theor. Comput. Sci.*, 410(17):1581–1588, 2009.

Claude d'Aspremont and Louis-Andr Grard-Varet. Incentives and incomplete information. *Journal of Public Economics*, pages 25–45, 1979.

Chrysanthos Dellarocas. Double marginalization in performance-based advertising: Implications and solutions. To appear in *Management Science*, 2012.

Nikhil R. Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *ACM Conference on Electronic Commerce*, pages 99–106, 2009.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1): 242–259, 2007.

Peter Ëso and Balazs Szentes. Optimal information disclosure in auctions and the handicap auction. *Review of Economic Studies*, 74(3):705–731, 2007.

Tomás Feder, Hamid Nazerzadeh, and Amin Saberi. Approximating nash equilibria using small-support strategies. In *ACM Conference on Electronic Commerce*, pages 352–354, 2007.

Jérémie Gallien. Dynamic mechanism design for online commerce. *Operations Research*, 54(2):291–310, 2006.

Alex Gershkov and Benny Moldovanu. Dynamic revenue maximization with heterogeneous objects: A mechanism design approach. *American Economic Journal: Microeconomics*, 1(2):168–198, 2009.

Andrew Goldberg, Jason Hartline, Anna Karlin, Mike Saks, and Andrew Wright. Competitive auctions. *Games and Economic Behavior*, 2006.

Google. http://www.google.com/intl/en/press/annc/payperaction.html, 2007.

Bryan Grow, Ben Elgin, and Moria Herbst. Click fraud: The dark side of online advertising. *BusinessWeek, Cover story*, 47(2-3), October 2 2006.

Faruk Gul and Andrew Postlewaite. Asymptotic efficiency in large exchange economies with asymmetric information. *Econometrica*, 60:1273–1292, 1992.

Mohammad Taghi Hajiaghayi, Robert D. Kleinberg, and David C. Parkes. Adaptive limited-supply online auctions. In *ACM Conference on Electronic Commerce*, pages 71–80, 2004.

Mohammad Taghi Hajiaghayi, Robert D. Kleinberg, Mohammad Mahdian, and David C. Parkes. Online auctions with re-usable goods. In *ACM Conference on Electronic Commerce*, pages 165–174, 2005.

J. Michael Harrison, N. Bora Keskin, and Assaf Zeevi (assaf@gsb.columbia.edu). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. To appear in *Management Science*, 145:106–123, 2010.

Nicole Immorlica, Kamal Jain, Mohammad Mahdian, and Kunal Talwar. Click fraud resistant methods for learning click-through rates. In *Internet and Network Economics, First International Workshop*, 2005.

Sham Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal dynamic mechanism design and the virtual pivot mechanism. *Working Paper*, 2010.

Fuhito Kojima and Mihai Manea. Incentives in the probabilistic serial mechanism. *Journal of Economic Theory*, 145:106–123, 2010.

Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *ACM Conference on Electronic Commerce*, pages 36–41, 2003.

Mohammad Mahdian and Kerem Tomak. Pay-per-action model for online advertising. In *Internet and Network Economics, Third International Workshop*, pages 549–557, 2007.

Dan Mitchell. Click fraud and halli-bloggers. *New York Times*, July 16 2005.

Roger Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th International Conference on World Wide Web*, pages 179–188, 2008.

Mallesh Pai and Rakesh Vohra. Optimal dynamic auctions and simple index rules. *Working paper*, 2009.

David Parkes. Online mechanisms. *Algorithmic Game Theory (Nisan et al. eds.)*, 2007.

David C. Parkes and Satinder P. Singh. An MDP-based approach for online mechanism design. In *Proceedings of the 17th Conference on Neural Information Processing Systems*, 2003.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: Incentive compatibility, profit maximization and information disclosure. *Working paper*, 2009.

Johan Roberts and Andrew Postlewaite. The incentives for price-taking behavior in large exchange. *Econometrica*, 44:115–129, 1976.

Maher Said. Auctions with dynamic populations: Efficiency and revenue maximization. To appear in *Journal of Economic Theory*, 2012.

James Schummer. Almost-dominant strategy implementation. *Games and Economic Behavior*, 48:154–170, 2004.

Andrzej Skrzypacz and Simon Board. Optimal dynamic auctions for durable goods: Posted prices and fire-sales. *Working paper*, 2010.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *21st Annual Conference on Learning Theory - COLT*, pages 343–354, 2008.

Stephen Spencer. Google deems pay-peraction as the "Holy Grail". *CNET News*, August 2007.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning, an introduction*. Cambridge: MIT Press/Bradford Books, 1998.

Gustavo Vulcano, Garrett van Ryzin, and Costis Maglaras. Optimal dynamic auctions for revenue management. *Management Science*, 48(11):1388–1407, 2002.

Robert Wilson. Game-theoretic approaches to trading processes. *Economic Theory: Fifth World Congress, ed. by T. Bewley*, pages 33–77, 1987.