Computational Game Theory **Spring Semester, 2009/2010**

Final project - Sponsored Search Auctions

Name: Vera Vsevolozhsky

1.1 Introduction

One of the more visible means by which the Internet has disrupted traditional activity is the manner in which advertising is sold. Offline, the price for advertising is typically set by negotiation or posted price. Online, much advertising is sold via auction. Most prominently,Web search engines like Google and Yahoo! auction space next to search results, a practice known as sponsored search. We consider how a search engine should select advertisements to display with search results, in order to maximize its revenue. Under the standard "$pay-per-click$" arrangement, revenue depends on how well the displayed advertisements appeal to users. The main difficulty stems from new advertisements whose degree of appeal has yet to be determined. Often the only reliable way of determining appeal is exploration via display to users, which detracts from exploitation of other advertisements known to have high appeal. Budget constraints and finite advertisement lifetimes make it necessary to explore as well as exploit.
In this summary we study the tradeoff between exploration and exploitation, modeling advertisement placement as a multi-armed bandit problem. Further, the traditional bandit formulations will be extended to account for budget constraints that occur in search engine advertising markets.
In addition multi-armed bandit mechanism truthfulness is considered and the price of truthfulness for "$pay-per-click$" auctions is derived.

1.1.1 The Advertisement Problem

Consider the following advertisement problem: there are m advertisers $y_1, y_2, ..., y_m$ who wish to advertise on a search engine. The search engine runs a large auction where each advertiser submits its bids to the search engine for the query phrases $q_1, q_2, ..., q_m$ in which it is interested. Advertiser y_i submits advertisement $a_{i,j}$ to target query phrase q_j , and promises to pay $b_{i,j}$ amount of money for each click on this advertisement, where $b_{i,j}$ is $y_i's$ bid for advertisement $a_{i,j}$. Given a user search query on phrase q_j, the search engine selects a constant number $C \geq 1$ of advertisements from the candidate set of advertisements $\{a_{*,j}\}$. The objective in selecting advertisements is to maximize the search engines total revenue. The arrival sequence of user queries is not known in advance. High revenue is achieved by displaying advertisements that have high bids as well as high likelihood of being clicked on

by users. Formally, the click-through rate (CTR), $c_{i,j}$, of advertisement $a_{i,j}$ is the probability of a user to click on advertisement $a_{i,j}$ given that the advertisement was displayed to the user for query phrase q_j .

1.1.2 Exploration/Exploitation Tradeoff

In this section we will discuss *exploration/exploitation* tradeoff. To maximize short-term revenue, the search engine should exploit its current, imperfect CTR estimates by displaying advertisements whose estimated $CTRs$ are large. On the other hand, to maximize long-term revenue, the search engine needs to explore, i.e., identify which advertisements have the largest $CTRs$. This kind of exploration entails displaying advertisements whose current CTR estimates are of low confidence, which inevitably leads to displaying some $low - CTR$ ads in the short-term. This kind of tradeoff between *exploration* and *exploitation* shows up often in practice and has been extensively studied in the context of the multi-armed bandit problem.

1.2 Multi-armed bandit problem

A multi-armed bandit problem is specifed by a set of strategies φ and a set Γ of possible *cost functions* on φ. Let $c_1, c_2, ..., c_T$ denote a sequence of *cost functions* chosen by an oblivious adversary. A multi-armed bandit algorithm is a randomized online algorithm for choosing a sequence of strategies $x_1, x_2, ..., x_T \in \varphi$ such that for all $t \geq 1$ the algorithm's choice of x_t depends only on its own random bits and on the values $c_1(x_1), c_2(x_2), ..., c_{t-1}(x_{t-1})$. The algorithm's objective is to minimize the average cost of the strategies it chooses. Its normalized regret is defied by the following formula:

$$Regret(algorithm, T) = max_{x \in \varphi} \frac{1}{T} E[\sum_1^T c_t(x_t) - c_t(x)].$$

We will first observe the application of multi-armed bandit algorithm to the advertisement problem, where using multi-armed bandit algorithm causes to the advertisers lying about their true bid values. And further we will review to the truthful mechanism analyzing.

1.2.1 Unbudgeted Unknown-CTR Advertisement Problem

In this section we will refer to the S. Pandey and C. Olston paper [3] and consider advertisement problem where budget(total amount of money advertiser y_i is willing to pay for the clicks on its advertisements in a day) constraints are absent and $CTRs$ are initially unknown. To solve the unbudgeted unknown-CTR advertisement problem, the authors create a multi-armed bandit problem instance for each query phrase q_j, where ads targeted for

the query phrase are the arms, bid values are the rewards and $CTRs$ are the payoff probabilities of the bandit instance. Let us assume $CTRs$ to be independent of one another. Since there are no budget constraints, each query phrase q_j is treated independently and each bandit instance is solved in isolation. The number of invocations for a bandit instance is not known in advance because the number of queries of phrase q_j in a given day is not known in advance. And as we already mentioned our goal is to determine a policy for activating the arms so as to maximize the total reward. Let us observe the policy suggested in [3]:

Policy MIX:
Let us define the priority $P_{i,j}$ of the ad $a_{i,j}$ as a function of its current CTR estimate $\tilde{c_{i,j}}$, its bid value $b_{i,j}$, the number of times it has been displayed so far $n_{i,j}$, and the number of times phrase q_j has been queried so far in the day, N_j, as follows:

$$P_{i,j} = (\tilde{c_{i,j}} + \sqrt{\frac{2ln(N_j)}{n_{i,j}}})b_{i,j}, \text{ for } n_{i,j} > 0, \text{ otherwise } P_{i,j} = \infty.$$

The policy MIX will act in the following way:
Initialization: Display each ad once.
Loop: For each query phrase q_j arrives at time t do:

1. Display all ads $i = \{1, 2, ..., C\}$, which has the highest priority $P_{i,j}$, targeted for query phrase q_j.
2. Update the CTR estimates $\tilde{c_{i,j}}$ as the average click-through rate observed so far, i.e., the number of times ad $a_{i,j}$ has been clicked on divided by the total number of times it has been displayed.

Policy MIX manages the exploration/exploitation tradeoff in the following way- the priority function, $P_{i,j}$, consists of two factors as follows:

1. Exploration factor, $\sqrt{\frac{2ln(N_j)}{n_{i,j}}}$, that becames smaller in time
2. Exploitation factor $\tilde{c_{i,j}}$.

Since $\tilde{c_{i,j}}$ can be estimated only when $n_{i,j} \geq 1$, the $P_{i,j}$ is set to ∞ for an ad which has never been displayed before. In [3] the authors proved that a performance bound for the number of policy's MIX mistakes for any $C > 0$ is $O(ln(n))$ on expectation, in n invocations. Since MIX policy construction is based on the policy $UCB1$ [2], the proof has been largely inherited from [2] (P. Auer, N. Cesa-Bianchi, and P. Fischer's paper). A mistake occurs when an ad $a_{i,j}$ that has priority less than some ad with the lowest priority, is mistakely displayed for

query phrase q_j. Thus, the authors bound the expected number of mistakes done by policy MIX by bounding the expected number of mistakes done by single mentioned ad $a_{i,j}$.
Let us consider the specific policy $UCB2$ from [2]. $UCB2$ is proposed under a slightly different reward model; we adapt it to our context to produce the following policy that we call $MIXR$. We replace the policy MIX mentioned in [3] by the *new* policy $MIXR$ and try to find the upper bound for the expected number of mistakes done by $MIXR$.

Policy MIXR:
Fist, we define the priority $P_{i,j}$ of the ad $a_{i,j}$ as a function of its current CTR estimate $\tilde{c_{i,j}}$, its bid value $b_{i,j}$, the number of times it has been displayed so far $n_{i,j}$(the number of epochs played by the arm i so far when query phrase q_j is run), and the number of times, N_j, query phrase q_j has been queried so far in the day as follows:
$P_{i,j} = (\tilde{c_{i,j}} + g(N_j, n_{i,j}))b_{i,j}$, where $g(n,r) = \sqrt{\frac{(1+\alpha)ln(\frac{en}{\tau(r)})}{2\tau(r)}}$, $0 < \alpha < 1$ and $\tau(r) = \lceil (1+\alpha)^r \rceil$, for $n_{i,j} > 0$, otherwise $P_{i,j} = \infty$.

The policy $MIXR$ will act in the following way:
Initialization: Set $n_{i,j} = 0$ for each $a_{i,j}$(arm (i,j)) and query phrase q_j. Display each ad once.
Loop: For each query phrase q_j arrives at time t do:

1. Select ads, $a_{i,j}$, where $i = \{1, 2, ..., C\}$, which has the highest priority $P_{i,j}$, targeted for query phrase q_j(=select arms maximizing $(\tilde{c_{i,j}} + g(N_j, n_{i,j}))b_{i,j}$,where $\tilde{c_{i,j}}$ is a reward obtained from playing $a_{i,j}$).

2. Display ads selected in 1(play arm (i,j)) exactly $\tau(n_{i,j}+1) - \tau(n_{i,j})$ times

3. Update the CTR estimates $\tilde{c_{i,j}}$ as the average click-through rate observed so far, i.e., the number of times ad $a_{i,j}$ has been clicked on divided by the total number of times it has been displayed.

4. Set $n_{i,j} \leftarrow n_{i,j} + 1$ (increase number of epochs by one for the selected $a_{i,j}$'s)

Note that similar to policy MIX, $MIXR$ has exploration/exploitation factor. And $P_{i,j} = \infty$ for $n < 1$. Next, we will observe a performance bound for the number of policy's $MIXR$ mistakes for any $C \geq 1$. Since $MIXR$ is adapted from $UCB2$ from [2], the proof of the following theorem is based on the proof of $UCB2$ performance bound from [2].

Theorem 1.1 *For any ad $a_{i,j} \in \{A_j - G_j\}, E[m_{i,j}(N_j)] = O(ln(N_j))$ where E denotes the expectation, A_j is the set of ads targeted query phrase q_j, G_j is the set of C ads of the highest priority expected rewards, $m_{i,j}(N_j)$ is the number of times ad $a_{i,j}$ is displayed by $MIXR$,*

N_j denotes the number of times query phrase q_j has been answered so far, assuming that $N_j \geq 1/(2\Delta_{i,j}^2)$ for all i,j where $\Delta_{i,j} = min_{a_{k,j} \in G_j}(c_{k,j}b_{k,j} - c_{i,j}b_{i,j})$.

Proof: Assume that $N_j \geq 1/(2\Delta_{i,j}^2)$ for all i, j and let $n_{i,j}^{\sim}$ be the largest integer such that $\tau(n_{i,j}^{\sim} - 1) \leq \frac{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}{2\Delta_{i,j}^2}$. Let us denote $L_j(N_j)$ to be the ad of the lowest priority value in G_j. Since mistake occures when $a_{i,j} \in \{A_j - G_j\}$ is selected to be displayed for q_i, we will consider only ads $a_{i,j} \in \{A_j - G_j\}$. Then after finish running $n_{i,j} - th$ epochs we have as follows:
$m_{i,j}(N_j) \leq 1 + \sum_{n_{i,j} \geq 1}(\tau(n_{i,j}) - \tau(n_{i,j} - 1)) \leq n_{i,j}^{\sim} + \sum_{n_{i,j} \geq n_{i,j}^{\sim}}(\tau(n_{i,j}) - \tau(n_{i,j} - 1))$.
$=>$ $\exists l \geq 0$, $\exists t \geq \tau(n_{i,j} - 1) + \tau(l)$, such that priority of some ad, $a_{i,j} \in \{A_j - G_j\}$ is $(c_{i,j}^{\sim} + g(t, n_{i,j} - 1))b_{i,j} \geq (c_{k,j}^{\sim} + g(t, l))b_{k,j}^{(\perp)}$, where $a_{k,j} = L_j(N_j) \in G_j$.
Let us observe the following two terms for $\gamma = c_{k,j}b_{k,j} - \alpha\Delta_{i,j}b_{k,j}/2$, where $c_{k,j}b_{k,j}$ is expected revenue for ad $a_{k,j}$:

1. $(c_{i,j}^{\sim} + g(N_j, n_{i,j} - 1))b_{i,j} \geq \gamma$, where $N_j \geq t$
2. $\exists i \geq 0$, $(c_{k,j}^{\sim} + g(t^{'}, i))b_{k,j} \leq \gamma$, where $t^{'} = \tau(n_{i,j} - 1) + \tau(i)$

Note that $g(t, r)$ is increasing in t and $N_j \geq t \geq t^{'}$. By contradiction, let us assume that none of these terms are true, then $\gamma > (c_{i,j}^{\sim} + g(N_j, n_{i,j} - 1))b_{i,j} \geq (c_{i,j}^{\sim} + g(t, n_{i,j} - 1))b_{i,j} \geq (c_{k,j}^{\sim} + g(t, i))b_{k,j} \geq (c_{k,j}^{\sim} + g(t^{'}, i))b_{k,j} > \gamma$. Then, condition $(\perp)$ can not be true. Hence, one of the terms must hold.

$=>$ $E[m_{i,j}(N_j)] \leq n_{i,j}^{\sim} + \sum_{n_{i,j} \geq n_{i,j}^{\sim}}(\tau(n_{i,j}) - \tau(n_{i,j} - 1))\ Pr[$display ad i from $\{A_j - G_j\}$ targeted $q_j] = n_{i,j}^{\sim} + \sum_{n_{i,j} \geq n_{i,j}^{\sim}}(\tau(n_{i,j}) - \tau(n_{i,j} - 1))\ Pr[$condition Y holds] {one of the above terms must be true} $= n_{i,j}^{\sim} + \sum_{n_{i,j} \geq n_{i,j}^{\sim}}(\tau(n_{i,j}) - \tau(n_{i,j} - 1))\ Pr[(c_{i,j}^{\sim} + g(N_j, n_{i,j} - 1))b_{i,j} \geq \gamma] + + \sum_{n_{i,j} \geq n_{i,j}^{\sim}} \sum_{i \geq 0}(\tau(n_{i,j}) - \tau(n_{i,j} - 1))\ Pr[(c_{k,j}^{\sim} + g(t^{'}, i))b_{k,j} \leq \gamma]$.

The assumption $N_j \geq 1/(2\Delta_{i,j}^2)$ implies $ln(2N_j\Delta_{i,j}^2) \geq 0$.
$=> ln(e) + ln(2N_j\Delta_{i,j}^2) \geq ln(e)$
$=> ln(2eN_j\Delta_{i,j}^2) \geq 1$.
$=>$ for any $n_{i,j} > n_{i,j}^{\sim}$, we obtain by definition of $n_{i,j}^{\sim}$ that

$$\tau(n_{i,j} - 1) > \frac{(1 + 4\alpha)ln(2eN_j\Delta_{i,j}^2)}{2\Delta_{i,j}^2} \tag{1.1}$$

and using(1.1)

$$g(N_j, n_{i,j} - 1) \leq \Delta_{i,j}\sqrt{\frac{(1+\alpha)ln(eN_j/\tau(n_{i,j} - 1))}{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}} \leq^{(*)} \Delta_{i,j}\sqrt{\frac{(1+\alpha)ln(2eN_j\Delta_{i,j}^2)}{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}} = \Delta_{i,j}\sqrt{\frac{1+\alpha}{1+4\alpha}}, \tag{1.2}$$

where $(*)$ holds due to (1.1) and $ln(2en_j\Delta_{i,j}^2) \geq 1$ ($\tau(n_{i,j}-1) > \frac{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}{2\Delta_{i,j}^2} > \frac{1}{2\Delta_{i,j}^2}$).

Recall Chernoff-Hoeffding bound: let $X_1, X_2, ..., X_n$ be random variables with common range [0 1] and such that $E[X_t|X_1, X_2, ..., X_{t-1}] = \mu$. Let $S_n = X_1 + X_2 + ... + X_n$. Then for all $a \geq 0$,

$$Pr[S_n \geq n\mu + a] \leq e^{-2a^2/n} \text{ and } Pr[S_n \leq n\mu - a] \leq e^{-2a^2/n} .$$

Using Chernoff-Hoeffding bound and result from (1.2) we get that
$Pr[(\tilde{c_{i,j}} + g(N_j, n_{i,j}-1))b_{i,j} \geq \gamma] =$
$Pr[(\tilde{c_{i,j}} + g(N_j, n_{i,j}-1))b_{i,j} \geq c_{i,j}b_{i,j} + \Delta_{i,j} - b_{k,j}\alpha\Delta_{i,j}/2]$ {by definition of $\Delta_{i,j}$} $\leq$
$e\{-2\tau(n_{i,j}-1)\Delta_{i,j}^2(1 - b_{k,j}\alpha/2 - b_{i,j}\sqrt{\frac{1+\alpha}{1+4\alpha}})^2\} \leq^{(**)}$
$e\{-2\tau(n_{i,j}-1)\Delta_{i,j}^2(1 - b_{k,j}\alpha/2 - b_{i,j}(1-\alpha))^2\} =$
$e\{-2\tau(n_{i,j}-1)\Delta_{i,j}^2(1 - b_{k,j}(1-\alpha/2))^2\} =$
$e\{-0.5\tau(n_{i,j}-1)\Delta_{i,j}^2(2 - b_{k,j}(2-\alpha/2))^2\}$.
Note that $(**)$ is correct because $\sqrt{\frac{1+\alpha}{1+4\alpha}} \leq (1-\alpha)$ for $0 < \alpha < 0.15$. It is easy to show. Assume by contradiction that it does not hold.
$=> \sqrt{\frac{1+\alpha}{1+4\alpha}} > (1-\alpha)$.
$=> \frac{1+\alpha}{1+4\alpha} > (1 - 2\alpha + \alpha^2)$.
$=> 1 + \alpha > 1 - 2\alpha + \alpha^2 + 4\alpha - 8\alpha^2 + 4\alpha^3$
$=> 1 - 7\alpha + 4\alpha^2 < 0$, that is impossible for such choice of α.

Note that $\tau(n_{i,j}) = \lceil(1+\alpha)^{n_{i,j}}\rceil \leq (1+\alpha)^{n_{i,j}} + 1 \leq \lceil(1+\alpha)^{n_{i,j}-1}\rceil(1+\alpha) + 1 =$
$\tau(n_{i,j}-1)(1+\alpha) + 1$. Then, $\tau(n_{i,j}) - 1 \leq \tau(n_{i,j}-1)(1+\alpha)$.
Let us define function $h(x) = (x-1)/(1+\alpha)$. Hence, for $\tau(n_{i,j}-1) \leq x \leq \tau(n_{i,j})$ and $n_{i,j} \geq 1$ we get that

$$h(x) \leq h(\tau(n_{i,j})) = \frac{\tau(n_{i,j})-1}{1+\alpha} \leq \frac{\tau(n_{i,j}-1)(1+\alpha)}{1+\alpha} = \tau(n_{i,j}-1)$$

$=> \sum_{n_{i,j} \geq \tilde{n_{i,j}}}(\tau(n_{i,j}) - \tau(n_{i,j}-1))\ Pr[(\tilde{c_{i,j}} + g(N_j, n_{i,j}-1))b_{i,j} \geq c_{i,j}b_{i,j} + \Delta_{i,j} - b_{k,j}\alpha\Delta_{i,j}/2] \leq$
$\leq \sum_{n_{i,j} \geq \tilde{n_{i,j}}}(\tau(n_{i,j}) - \tau(n_{i,j}-1))e^{-\tau(n_{i,j}-1)c} \leq \int_0^\infty e^{-ch(x)}dx$, where $c = \Delta_{i,j}^2(2 - b_{k,j}(2-\alpha/2))^2$
But, $\int_0^\infty e^{-ch(x)}dx =$ {by def. of $h(x)$} $\int_0^\infty e^{-c(x-1)/(1+\alpha)}dx = \frac{1+\alpha}{c}e^{\frac{c}{1+\alpha}} \leq \frac{1+\alpha}{c}e$, for $\frac{c}{1+\alpha} < 1$.

Next, using Chernoff-Hoeffding bound and def. of t' we obtain as follows:
$Pr[(\tilde{c_{k,j}} + g(t', i))b_{k,j} \leq c_{k,j}b_{k,j} - \alpha\Delta_{i,j}b_{k,j}/2] = Pr[(\tilde{c_{k,j}} + g(t', i)) \leq c_{k,j} - \alpha\Delta_{i,j}/2] \leq$
$e\{-\tau(i)(\alpha\Delta_{i,j})^2/2 - (1+\alpha)ln(\frac{e\tau(n_{i,j}-1)+\tau(i)}{\tau(i)})\}$.

Hence, $\sum_{n_{i,j} \geq \tilde{n_{i,j}}}\sum_{i \geq 0}(\tau(n_{i,j}) - \tau(n_{i,j}-1))\ Pr[(\tilde{c_{k,j}} + g(t', i)) \leq c_{k,j} - \alpha\Delta_{i,j}/2] \leq$

$\sum_{i\geq 0}\sum_{n_{i,j}\geq \tilde{n_{i,j}}}(\tau(n_{i,j})-\tau(n_{i,j}-1))e\{-\tau(i)(\alpha\Delta_{i,j})^2/2-(1+\alpha)ln(\frac{e\tau(n_{i,j}-1)+\tau(i)}{\tau(i)})\}^{(***)}$

Since $e^{-u-w}=e^{-u}*e^{-w}$, we get that
$e\{-\tau(i)(\alpha\Delta_{i,j})^2/2-(1+\alpha)ln(\frac{e\tau(n_{i,j}-1)+\tau(i)}{\tau(i)})\}=e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}*e\{-(1+\alpha)ln(\frac{e\tau(n_{i,j}-1)+\tau(i)}{\tau(i)})\}=$
$e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}*e\{-(1+\alpha)(1+ln(\frac{\tau(n_{i,j}-1)}{\tau(i)}+1))\}\leq$
$e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}*e\{-(1+\alpha)ln(\frac{\tau(n_{i,j}-1)}{\tau(i)}+1)\}=$
$e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}(\frac{\tau(n_{i,j}-1)}{\tau(i)}+1)^{-(1+\alpha)}$. Note that $\tau(i)\geq 1$. Recall $h(x)\leq\tau(n_{i,j}-1)$.
$=>(***)\leq\sum_{i\geq 0}e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}\sum_{n_{i,j}\geq\tilde{n_{i,j}}}(\tau(n_{i,j})-\tau(n_{i,j}-1))(\frac{\tau(n_{i,j}-1)}{\tau(i)}+1)^{-(1+\alpha)}\leq$
$\sum_{i\geq 0}e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}\int_0^\infty(1+\frac{h(x)}{\tau(i)})^{-(1+\alpha)}dx=$
$\sum_{i\geq 0}\tau(i)e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}(\frac{1+\alpha}{\alpha})(1-\frac{1}{\tau(i)(1+\alpha)})^{-\alpha}\leq$
$(\frac{1+\alpha}{\alpha})^{(1+\alpha)}\sum_{i\geq 0}\tau(i)e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}$.

Let us observe $\sum_{i\geq 0}\tau(i)e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}$. For $(1+\alpha)^{(x-1)}\leq\tau(i)\leq(1+\alpha)^x+1$ and $i\leq x\leq i+1$ we get as follows:

$$\sum_{i\geq 0}\tau(i)e\{-\tau(i)(\alpha\Delta_{i,j})^2/2\}\leq 1+\int_1^\infty((1+\alpha)^x+1)e\{-0.5(1+\alpha)^{x-1}(\alpha\Delta_{i,j})^2\}dx \qquad (1.3)$$

Let us denote $v=(1+\alpha)^x$ and change integration variable in (1.3). Hence, we obtain that $(1.3)\leq 1+\int_1^\infty\frac{v+1}{vln(1+\alpha)}e^{-\frac{v(\alpha\Delta_{i,j})^2}{2(1+\alpha)}}dv=1+\frac{1}{ln(1+\alpha)}[\frac{e^{-\lambda}}{\lambda}+\int_\lambda^\infty e^{-x}/xdx]$, where $\lambda=\frac{(\alpha\Delta_{i,j})^2}{2(1+\alpha)}$.
For $\alpha>0$ and $\Delta_{i,j}<1$ we obtain that $0<\lambda<0.25$. Let us denote $F(\lambda)$ to be $[\frac{e^{-\lambda}}{\lambda}+\int_\lambda^\infty e^{-x}/xdx]$.
Hence, $F^{'}(\lambda)=-2e^{-\lambda}+\int_\lambda^\infty e^{-x}/xdx$ and $F^{''}(\lambda)=2\lambda e^{-\lambda}-\int_\lambda^\infty e^{-x}/xdx$. Let us find the extremum points of $F(\lambda)$. $F^{'}(\lambda)=0$ at $\lambda=0.0108$, when we consider $0<\lambda<0.25$. The second derivative,$F^{''}(\lambda)<0$, for such choice of λ. Hence, F has its maximum(unique) point in $\lambda=0.0108$, that is $F(0.0108)<11/10$.
$=>F(\lambda)<11/(10\lambda)$
$=>1+\frac{1}{ln(1+\alpha)}F(\lambda)<1+\frac{1}{ln(1+\alpha)}\frac{11}{10\lambda}$

Combining everything together, we get that
$E[m_{i,j}(N_j)]\leq\tau(\tilde{n_{i,j}})+\frac{(1+\alpha)e}{c}+(\frac{1+\alpha}{\alpha})^{1+\alpha}[1+\frac{11}{10\lambda ln(1+\alpha)}]$, where $c=\Delta_{i,j}^2(2-b_{k,j}(2-\alpha/2))^2$ and $\lambda=\frac{(\alpha\Delta_{i,j})^2}{2(1+\alpha)}$. Recall that we have already shown that $\tau(n_{i,j})\leq\tau(n_{i,j}-1)(1+\alpha)+1$ and for any $n_{i,j}>\tilde{n_{i,j}}$, we obtain by definition of $\tilde{n_{i,j}}$ that $\tau(n_{i,j}-1)>\frac{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}{2\Delta_{i,j}^2}$. Note that $\tilde{n_{i,j}}\geq 1$.
Hence, $E[m_{i,j}(N_j)]\leq\tau(n_{i,j}-1)(1+\alpha)+1+\frac{(1+\alpha)e}{c}+(\frac{1+\alpha}{\alpha})^{1+\alpha}[1+\frac{11}{10\lambda ln(1+\alpha)}]\leq$
$\frac{(1+4\alpha)ln(2eN_j\Delta_{i,j}^2)}{2\Delta_{i,j}^2}(1+\alpha)+1+\frac{(1+\alpha)e}{c}+(\frac{1+\alpha}{\alpha})^{1+\alpha}[1+\frac{11}{10\lambda ln(1+\alpha)}]$.

Hence, $E[m_{i,j}(N_j)] = O(ln(N_j))$.

For $MIXR$ policy we obtain that $E[m_{i,j}(N_j)] \leq \frac{(1+4\alpha)(1+\alpha)ln(N_j)}{2\Delta_{i,j}^2}$ plus constant.
Then, there is exists such α so that leading constant, $\frac{(1+4\alpha)(1+\alpha)}{2\Delta_{i,j}^2}$, could be arbitary close to $\frac{1}{2\Delta_{i,j}^2}$. For example, we can choose α to be 0.001. This result is better than the bound obtained for the policy MIX in [3]. For policy MIX, we get that $E[m_{i,j}(N_j)] \leq \frac{8b_{i,j}^2 ln(N_j)}{\Delta_{i,j}^2}$ plus constant, where the leading constant is $\frac{8b_{i,j}^2}{\Delta_{i,j}^2}$.

□

Given the above result we obtain that expected number of mistakes made by $MIXR$ for a set of the query phrases Q is

$$\sum_{q_j \in Q} \sum_{a_{i,j} \in \{A_j - G_j\}} E[m_{i,j}(N_j)] = O(lnN).$$

1.2.2 Budgeted Unknown-CTR Advertisement Problem

We now turn to the more challenging case in which advertisers can specify daily budgets. In the previous section, in the absence of budget constraints, we were able to treat the bandit instance created for a query phrase independent of the other bandit instances. However, budget constraints create dependencies between query phrases targeted by an advertiser. To model this situation, a new kind of bandit problem that is called Budgeted Multi-armed Multi-bandit Problem (BMMP), in which multiple bandit instances are run in parallel under overarching budget constraints has been suggested by authors in [3]. The policy, $BPOL$, that suggested in [3] as follows:

1. Run $|B|$ instances of POL in parallel, denoted $POL_1, POL_2, ...POL_{|B|}$.

2. Whenever bandit instance B_i is invoked:
   2.1 Discard any arm(s) of B_i whose type's budget is newly depleted, i.e., has become depleted since the last invocation of B_i.
   2.2 If one or more arms of B_i was discarded during step 1, restart POL_i.
   2.3 Let POL_i decide which of the remaining arms of B_i to activate

In the [3] it is proved that $bpol(N) \geq opt(N)/2 - O(f(N))$ for any N, where $bpol(N)$ and $opt(N)$ denote the total expected reward obtained after N invocations by $BPOL$ and OPT, respectively, and $f(n)$ denotes the expected number of mistakes made by POL after n invocations of the the regular multi-armed bandit problem. Here we give a high-level outline

of the proof. For simplicity the authors focus on the case, where $C = 1$. Since bandit arms generate rewards stochastically, it is not clear how we should compare $BPOL$ and OPT. For example, even if $BPOL$ and OPT behave in exactly the same way (activate the same arm on each bandit invocation), there is no guarantee that both will have the same total reward in the end. To enable meaningful comparison, a payoff instance, I, is define, such that $I(i, n)$ denotes the reward generated by arm i of bandit instance $S(n)$ for invocation n in payoff instance I. The outcome of running $BPOL$ or OPT on a given payoff instance is deterministic because the rewards are fixed in the payoff instance. Hence, we can compare $BPOL$ and OPT on per payoff instance basis. Since each payoff instance arises with a certain probability, $P(I)$, by taking expectation over all possible payoff instances of execution we can compare the expected performance of $BPOL$ and OPT. Next, three categories of invocation n are observed:

1. The arm activated by OPT, $O(I, n)$, is of smaller or equal expected reward in comparison to the arm activated by $BPOL$, $B(I, n)$. The expected reward of an arm is the product of its payoff probability and reward.
2. Arm $O(I, n)$ is of greater expected reward than $B(I, n)$, but $O(I, n)$ is not available for $BPOL$ to activate at invocation n due to budget restrictions.
3. Arm $O(I, n)$ is of greater expected reward than $B(I, n)$ and both arms $O(I, n)$ and $B(I, n)$ are available for $BPOL$ to activate, but $BPOL$ prefers to activate $B(I, n)$ over $O(I, n)$.

Next, the authors prove that

$$bpol_k(N) = \sum_I (P(I) \sum n \in_{N^k(I)} I(B(I, n), n)), \text{ where } k = 1, 2, 3$$

and similar for $opt_k(N)$ Then the proof of each of the following bounds are provided:

1. $opt_1(N) \leq bpol_1(N)$
2. $opt_2(N) \leq bpol(N) + (|T|\tau_{max})$, where $|T|$ denotes the number of arm types and τ_{max} denotes the maximum reward. In $BMMP$ each arm has an associated *type*. Each type $T_i \in T$ has budget $d_i \in [0, \infty]$, which specifies the maximum amount of reward that can be generated by activating all arms of that type. Once the specified budget is reached for a type, no further reward will be earned from activating arms of that type.
3. $opt_3(N) = O(f(N))$

And finally, the following theorem is proved:

Theorem 1.2 *$bpol(N) \geq opt(N)/2 - O(f(N))$ for any N.*

Proof: $opt(N) = opt_1(N)+opt_2(N)+opt_3(N) \leq bpol_1(N)+bpol(N)+(|T|\tau_{max})+O(f(N))$ {by above three bounds} $\leq 2bpol(N) + O(f(N))$. Hence, $bpol(N) \geq opt(N)/2 - O(f(N))$ for any N. □

Note that $f(N)$ depends on the policy we work with (for example, for the policy MIX we obtain that $f(N) = O(logN)$).

Next, we derive $BMIXR$ policy (similar to $BMIX$ policy derivation suggested in [3]).

Policy BMIXR:

Loop: for each query phrase q_j arrives:

1. For ads whose advertisers have not depleted their budgets yet, compute the priorities and display the C ads as defined in $MIXR$.

2. Update the CTR estimates $\tilde{c_{i,j}}$ of the displayed ads by monitoring user clicks.

It is easy to see that $BMIXR$ is the instance of $BMMP$, while using $MIXR$ as an input policy. Note that it is not necessary to restart the $MIXR$ instance for q_j when an advertisers budget is depleted as done in the generic $BPOL$, since $MIXR$ has state on per add basis, so it can continue from where it left off if some ads are removed.

In the previous section we showed that for $MIXR$ policy we obtained that $f(N)$ is $O(lnN)$ for any $C \geq 1$. Hence, from Theorem 1.2, we conclude that the average revenue generated by $BMIXR$ is at least $opt(N)/2-O(lnN)$ for any $C \geq 1$, where $opt(N)$ denotes the optimal revenue generated from answering N user queries.

1.2.3 Conclusions

From [3] we studied how a search engine should select which ads to display in order to maximize revenue, when click-through rates are not initially known. We dealt with the underlying exploration/exploitation tradeoff using multi-armed bandit theory. The main contribution of S.Pandey and C. Olston in [3] was to bandit theory by proposing a new variant of the bandit problem that we call budgeted multi-armed multi-bandit problem ($BMMP$). A policy was proposed for solving $BMMP$ and derived a performance guarantee.

We consider a specific policy from [2] called $UCB2$, and policy MIX from [3] and derived new policy called $MIXR$. $UCB2$ from [2] was proposed under a slightly different reward model; we adapt it to our context to produce the new policy $MIXR$. We showed a performance bound of $O(ln(N))$ mistakes for $MIXR$ for any $C \geq 1$ and used $MIXR$ as the instance of $BMMP$ suggested in [3]. Similar to $BMIX$ the obtained policy $BMIXR$ has the average revenue that is at least $opt(N)/2 - O(ln(N))$ for any $C \geq 1$.

But, the main question is multi-armed bandit mechanism truthful, as it was suggested? The answer is no. Each advertiser acts as utility-maximizing player, the payments charged by mechanism depends on his bid declaration/clicks, which are not charged according to the second maximum price, for example. This cause to advertiser avoids declaring his real value in order to pay less per click. On the other hand, this manipulation value will affect auction's revenue, which depends on advertisers bid value that could be not truthful. Thus, we are interested in designing mechanisms which are truthful (in dominant strategies), and in the next section, we will analyze results obtained by Devanur and Kakade in their work([4]), where the truthful auctions are considered.

1.2.4 Characterizing Truthful Pay-Per-Click Auctions

Main contribution:

Devanur and Kakade in [4] have analyzed the problem of designing a truthful pay-per-click auction where the $CTRs$ of the bidders are unknown to the auction. They studied the truthful MAB mechanism with focus on maximization the revenue. For the truthfulness of the mechanism, they considered the expected Vickrey auction revenue that is hoped to be achieved by knowing the true $CTRs$ (this is actually the OPT that could be reached by auction). And truthful regret is defined to be the dfference between the expected revenue of the played auction and this Vickrey revenue. The main contribution of the authors in [4] is showing the existence of an algorithm with sublinear (in T) truthful regret by proving the upper bound on truthful regret, that is $O(b_{max}n^{1/3}T^{2/3}\sqrt{log(nT)})$. And proving that any truthful mechanism must have truthful regret $\Omega(T^{2/3})$, that is the lower bound. Further, the lower bound is extended to the case where the bidders submit a single bid value at the start, and the mechanism only charges at the end of the T rounds. The lower bound that is proved in the latest case is $\Omega(T^{2/3})$.

The model for a single-slot pay-per-click auction:

- T time steps(rounds) of the repeated auction would be played
- n advertisers, each of whom values a 'click', while the auction can only assign 'impressions'
- b_i^t - advertiser $i's$ bid for click in some round $t = 1, 2, ..., T$
- v_i^t - advertiser $i's$ true value for a click at time $t = 1, 2, ..., T$, that is his private information

- x^t - allocation vector, where $x_i^t = 1$ iff the slot allocation is to advertiser i at time t, and x_j^t is zero for all $j \neq i$
- c_i^t - event which is is equal to 1 if the item was clicked on and 0 otherwise (c_i^t is observed iff $x_i^t = 1$)
- p_i^t - payment that is charged the advertiser i at the end of the round t(charging is done by auction iif $c_i^t = 1$)
- $A = \sum_{i,t} p_i^t$-the revenue of the auction
- ρ_i - click probability (CTR), that is obtained when i.i.d event, $c_i^t = 1$

Proof ideas and results:

In this section we will give a brief explanation and key ideas to the proof of upper/lower bounds as it mentioned in [4], sometimes omitting technical details of the proofs.

1. **Upper bound analysis:** The algorithm is quite simple. For the first τ steps, the auction explores. By this we mean that the algorithm allocates the item to each bidder for $\lfloor \tau/n \rfloor$ steps (and it does so non-adaptively in some deterministic order). All prices are zero during this exploration phase. After this exploration phase is over($t \geq \tau$),the algorithm learned $\tilde{\rho}_i$ - the empirical estimate of the CTR, where for all advertisers $i = \{1, 2\}$, with probability of $1 - \delta$, holds that

$$\rho_i \leq \tilde{\rho}_i + \sqrt{2\lfloor n/\tau \rfloor log(n/\delta)} = \tilde{\rho}_i{}^+$$

   Next, from some point of time $t \geq \tau$, mechanism allocates slot for advertiser i^*, such that $i^* = argmax_i \tilde{\rho}_i{}^+ b_i^t$. and the price charged from advertiser is $p_i^t = \frac{smax_i \tilde{\rho}_i{}^+ b_i^t}{\tilde{\rho}_{i^*}^+}$, where the operator $smax$ means to charge the second maximum value. In the paper the authors provide proof of truthfulness of the described mechanism by considering a set of positive weights w_i and showing that we could construct truthful auction with this vector w_i in the following manner: let winner at time t be $i^* = argmax_i w_i b_i^t$ and charge him the amount $p_i^t = \frac{smax_i w_i b_i^t}{w_{i^*}}$. Instead of weight w_i we use p_i^+ that is not a function of bid. And then prove that $T - Regret = O(b_{max} n^{1/3} T^{2/3} \sqrt{log(nT)})$ for $\delta = 1/T$ and $\tau = n^{1/3} T^{2/3} \sqrt{log(nT)}$ by bounding $smax_i p_i b_i^t - \frac{smax_i \tilde{p_i^+} b_i^t}{\tilde{p_i^+}} p_i^*$ term, that is the difference between OPT revenue could've been obtained by the mechanism and expected revenue obtained by mechanism at time t, and summing over all $t's$. Note that for the multi-armed bandit mechanism, such algorithms typically also achieve a regret of the same order.

2. **Lower bound analysis:** The proof technique shows that any pay-per-click auction must have the property that it behaves as an 'explore/exploit' algorithm, where when it explores, it must charge zero, and when it exploits, it cannot use this information for setting future prices. The proof techniques go through the results on truthful pricing (Myerson [1981], Hartline and Karlin [2007]), which (generally) characterize how to truthfully price any allocation scheme. The additional constraint the authors use on this truthful pricing scheme is an informational one - the auction must only use information from the observed allocations. Note that Gonen and Pavlov consider the same problem in their work[6]. But their auction is not truthful, since for the allocation given by their auction, there is a unique pricing that would make it truthful, but this price depends on clicks that are not observed by the auction, which is lower bound technique in [4] imply. The technique shows how to obtain restrictions on the pricing scheme, based on both truthfulness and bandit feedback. Now, let us observe the proof as follows:

   In the proof of the lower bound the authors make two assumptions as follows:

   (a) auction must be $scale - invariant$ i.e., for all $\lambda > 0$, $x(b) = x(\lambda b)$

   (b) auction must be $non - degenerate$, i.e., if for all bids b_i^t there exists a sufficiently small interval I of positive length containing b_i^t such that for all other bids, clicks and time $t^{'}$, replacing b_i^t with any $b \in I$ does not change $x_i^{t^{'}}$.

   The model where an advertiser submits a bid for each time step is considered.
   First, the authors characterize the restriction imposed on the allocation function by truthfulness. They used Myerson theorem for characterizing truthful auctions, where each bidder submits a single bid. Since the advertiser's value could remain the same over all time periods(note that in our case each advertiser submits his bid for each time step) and one strategy he could take is to submit the same bid (which could be different from his true value) for all time steps, the Myerson theorem characterization still holds, and then apply it to the cumulative prices charged over the auction, leading to the following pricing restriction:

   Theorem 1.3 *For a fixed click sequence and $y_i = \sum_t x_i^t c_i^t$ and $p_i = \sum_t p_i^t$, if an auction x is truthful then y_i is monotonically increasing in b_i and the price, $p_i(b)$, charged to i is exactly $b_i y_i(b) - \int_{z=0}^{b_i} y_i(z, b_{-i})dz$.*

   Using this theorem it could be proved that every round of the auction is truthful.
   It is straightforward to see that the truthful pricing rule also implies that these must

be the instantaneous prices, and that instantaneously, the x_i^t(and so y_i^t) must be monotonic in b_i^t. To see this, we consider the case where the current round is effectively the advertiser's last round. Then, advertiser's values for the remaining rounds will be zero. Then, every round of the auctions is truthful. Hence, the whole mechanism is truthful.
=> The allocation function has to be such that the prices can always be calculated exactly. The further proof shows that the allocation function only has functional dependence on the clicks observed during certain time periods that are $'non-competitive'$. Note that we observe the case where there are two advertisers. Next, the following lemma is proved:

Lemma 1.4 *If τ is competitive w.r.t advertiser 1, i.e., for all b_2, there exist b_1,so that $x_1^\tau(b_1, b_2) = 1$, then x_1^t does not depend on c_2^τ, i.e.,there is no b_1 and b_2 such that $x_1^t(b_1, b_2, c_2^\tau) \neq x_1^t(b_1, b_2, 1 - c_2^\tau)$.*

Now we observe that in order for x_1^t to have a functional dependence on c_2^τ, the auction must observe c_2^τ. But in this case $x_2^\tau(b_1, b_2) = 1$, which means that a slot is given to advertiser 2 at time τ. The main idea that is used in the proof is to use assumption that the time τ is competitive w.r.t advertiser 1, and x_1^t does depend on c_2^τ, by contradiction, and to use the observation above. Note that in this is the place where we use our assumption of the non-degeneracy. Next, the following corollary is proved:

Corollary 1.5 *If τ is not competitive w.r.t advertiser 1, then $p_1^\tau = p_2^\tau = 0$*

And finally, the main theorem is proved:

Theorem 1.6 *For every non−degenerate, scale−invariant and always truthful pay-per-click auction (with 2 advertisers), there exists a set of bids bounded in [0; 1] and ρ_i such that T-Regret $= \Omega(T^{2/3})$.*

Because of Corollary it is enough to prove that given that the number of non-competitive rounds is $o(T^{2/3})$ with prob. $1 - o(1)$, the $T - Regret = \Omega(T^{2/3})$. This is because the auction does not profit in the non-competitive time τ and the expected revenue of the auction would be zero for the $o(T^{2/3})$ rounds. Thus we will always have regret, that is $\Omega(T^{2/3})$.
Next, the authors have shown that $T - Regret = OPT - E[p_1 + p_2] = \Omega(T^{2/3})$, where p_i defined as in the mentioned **truthful pricing rule**.
=> It is enough to show that $T - Regret = OPT - E[y_1 b_1 + y_2 b_2] = OPT - E_C[\rho_1 b_1 x_1 + \rho_1 b_1 x_1] = \Omega(T^{2/3})$ (since integral in **the truthful pricing rule** is positive and $E_{c_2^t}[y_2^t | c_2^1 ... c_2^{t-1}] = \rho_2 x_2^t$) implying $E_C[y_2^t] = \rho_2 E_C[x_2^t]$.

In order to proof that $T - Regret = \Omega(T^{2/3})$, two instances are considered as follows:
1. Instance 1: $(\rho_1, b_1) = (1, 0.5)$ and $(\rho_2, b_2) = (0.5 + \delta, 1)$
2. Instance 2: $(\rho_1, b_1) = (1, 0.5)$ and $(\rho_2, b_2) = (0.5 - \delta, 1)$
where $\delta = T^{-1/3}$.
Next the following claim is proved:

Claim 1.7 *For all click sequences C, if $x_1(0.5, 1, C) \geq T/2$ (resp. $x_1(0.5, 1, C) \leq T/2$) then the loss for C is at least $\delta T/2$ for Instance 1 (resp. Instance 2).*

and let χ be a function of C that is 1 if the loss for that click sequence is $\geq \delta T/2$ for Instance 1, and is 0 otherwise (loss is $\geq \delta T/2$ for Instance 2, as guaranteed by above Claim). Note that χ only depends on x_1, which only depends on on the clicks in the non-competitive rounds. Hence, χ can be represented as a boolean decision tree of depth $n = o(\frac{1}{\delta^2})$.

Next, a technical lemma is proved, which shows that such a function can essentially not distinguish between the two instances.

Lemma 1.8 *Let P_1 and P_2 be probability distributions on $\{0,1\}^T$ generated by i.i.d samples w.p $0.5+\delta$ and $0.5-\delta$ respectively. Then for all functions $\chi : \{0,1\}^T \to \{0,1\}$ that can be represented as decision trees of depth $o(1/\delta^2)$ either $\sum_{c\in\{0,1\}^T} P_1(c)\chi(c)$ or $\sum_{c\in\{0,1\}^T} P_2(c)(1-\chi(c))$ is $\Omega(1)$.*

Thus, we can apply Lemma 1.8 to χ. From Lemma 1.8, either $\sum_{c\in\{0,1\}} P_1(c)\chi(c) = 1$ or $\sum_{c\in\{0,1\}} P_2(c)(1-\chi(c))$ is $\Omega(1) = 1$. If the former holds, this says that the probability that the loss is $\Omega(\delta T) = \Omega(T^{2/3})$ for Instance 1 is $\Omega(1) = 1$. Thus the expected loss is $\Omega(T^{2/3})$. The other case implies an expected loss of $\Omega(T^{2/3})$ on Instance 2. Hence, $T - Regret = \Omega(T^{2/3})$.
Independently and concurrently, M. Babaioff, Y. Sharma, A. Slivkins have worked on the same problem and presented results in [5] that analogous to above lower bound proof.

References:

[1] Vijay V. Vazirani; Nisan, Noam; Tim Roughgarden; Eva Tardos (2007). Algorithmic Game Theory. Cambridge, UK: Cambridge University

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multi-Armed Bandit Problem. Machine Learning,47:235256, 2002.

[3] Sandeep Pandey, Christopher Olston. Handling advertisements of unknown quality in search advertising (2006)

[4] Nikhil R. Devanur, Sham M. Kakade, The Price of Truthfulness for Pay-Per-Click Auctions, 2009

[5] M. Babaioff, Y. Sharma, A. Slivkins, Characterizing Truthful Multi-Armed Bandit Mechanisms (2008)

[6] Rica Gonen,Elan Pavlov, An Incentive-Compatible Multi-Armed Bandit Mechanism(2007)

[7] Robert Kleinberg, Notes from Week 8: Multi-Armed Bandit Problems(Spring 2007, CS 683- Learning, Games, and Electronic Markets)