Showing Relevant Ads via Context Multi-Armed Bandits

Tyler Lu Dávid Pál
{ttlu,dpal}@cs.uwaterloo.ca
David R. Cheriton School of Computer Science
University of Waterloo, ON, Canada

Martin Pál
mpal@google.com
Google, Inc.
New York, NY

Abstract

We study *context* multi-armed bandit problems where the context comes from a metric space and the payoff satisfies a Lipschitz condition with respect to the metric. Abstractly, a context multi-armed bandit problem models a situation where, in a sequence of independent trials, an online algorithm chooses an action based on a given *context* (side information) from a set of possible actions so as to maximize the total payoff of the chosen actions. The payoff depends on both the action chosen and the context. In contrast, context-*free* multi-armed bandit problems, a focus of much previous research, model situations where no side information is available and the payoff depends only on the action chosen.

Our problem is motivated by sponsored web search, where the task is to display ads to a user of an Internet search engine based on her search query (context) so as to maximize the click-through rate of the ads displayed. We cast this problem as a context multi-armed bandit problem where queries and ads form metric spaces and the payoff function is Lipschitz with respect to both the metrics. For any $\epsilon > 0$ we present an algorithm with regret $O(T^{\frac{a+b+1}{a+b+2}+\epsilon})$ where a, b are the covering dimensions of the query space and the ad space respectively. We prove a lower bound $\Omega(T^{\frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+2}-\epsilon})$ for the regret of any algorithm where $\tilde{a}, \tilde{b}$ are packing dimensions of the query spaces and the ad space respectively. For finite spaces or bounded subsets of Euclidean spaces, this gives an (almost) matching upper and lower bound.

1 Introduction

Internet search engines, such as Google, Yahoo! and MSN, receive revenue from advertisements shown to a user's query. Whenever a user decides to click on an ad displayed for a search query, the advertiser pays the search engine. Thus, part of the search engine's goal is to display ads that are most relevant to the user in the hopes of increasing the chance of a click, and possibly increasing its expected revenue. In order to achieve this, the search engine has to learn over time which ads are the most relevant to display for different queries. On the one hand, it is important to *exploit* currently relevant ads, and on the other hand, one should *explore* potentially relevant ads.

This problem can be naturally posed as a multi-armed bandit problem with *context*. Here by context we mean a user's query. Each time a query x arrives and an ad y is displayed there is an (unknown) probability $\mu(x, y)$ that the user clicks on the ad.[1] We call $\mu(x, y)$ the click-through rate of x and y.

Our goal is to design an online algorithm, which given a query in each time step and a history of past queries and ad clicks, displays an ad to maximize the expected number of clicks. In our setting, we make a crucial yet very natural assumption that the space of queries and ads are endowed with a metric and the click-through rate $\mu(x, y)$ satisfies a Lipschitz condition with respect to each coordinate. Informally, we assume that the click-through rates of two similar ads for the same query are close, and that of two similar queries for the same ad are also close. Lastly, we assume that the queries are fixed in advance by an adversary and revealed in each time step (aka *oblivious* adversary).

Clearly, the best possible algorithm—*Bayes optimal* —displays, for a given query, the ad which has the highest click-through rate. Of course, in order to execute it the click-through rates must be known. Instead we are interested in algorithms that do not depend on the knowledge the click-through rates and whose performance is still asymptotically the same as that of the Bayes optimal. More precisely, for any algorithm A, we consider the expected difference between the number of clicks that the Bayes optimal receives and A receives for T queries. This difference is called the *regret* of A and is denoted by $\mathcal{R}_A(T)$. An algorithm is said to be asymptotically Bayes optimal if the *per-query* regret $\mathcal{R}_A(T)/T$ approaches 0 as $T \to \infty$ for any sequence of queries. The algorithm we present in this paper has this property.

The standard measure of quality of an asymptotically Bayes optimal algorithm is the speed of convergence at which per-round regret approaches zero. Equivalently, one measures the growth of the regret $\mathcal{R}_A(T)$ as $T \to \infty$. The bounds are usually of the form $\mathcal{R}_A(T) = O(T^\gamma)$ for some $\gamma < 1$. Such regret bounds are the standard way of measuring performance of algorithms for multi-armed bandit problems, for online learning problems and, more broadly, for reinforcement learning problems.

Our contribution: The main contribution of this paper are upper and lower bounds on the regret, independent of the click-through rates, in terms of the covering and packing dimensions of the query space and the ad space, respectively. The covering dimension of a metric space is defined as the smallest d such that the number of balls of radius r required to cover the space is $O((1/r)^d)$. The packing dimension, is defined as the largest $\tilde{d}$ such that there for any r there exists a subset of disjoint balls of radius r of size $\Omega((1/r)^{\tilde{d}})$.

For the upper bound, we present an algorithm, which we call the query-ad-clustering algorithm, that for any $\epsilon > 0$ achieves regret at most $O(T^{\frac{a+b+1}{a+b+2}+\epsilon})$ where a, b are the covering dimensions of the query space and the ad space, respectively. We present lower bound $\Omega(T^{\frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+2}-\epsilon})$ on regret of any algorithm on a problem where $\tilde{a}, \tilde{b}$ are the packing dimensions of the query and ads space respectively. More precisely, our results are stated in the following theorem.

[1] For simplicity we assume that only one ad is displayed per query. Actual search engines usually display multiple ads at once.

Theorem 1. *Consider a context Lipschitz multi-armed bandit problem with query space X and ads space Y of size at least 2. Let a, b be the covering dimensions of X, Y respectively. Let $\tilde{a}, \tilde{b}$ be the packing dimensions of X, Y respectively. Then,*

- *For any $\gamma > \frac{a+b+1}{a+b+2}$, there exists an algorithm A and positive constants T_0, C such that for any instance (i.e. click-through rates) μ, any $T \geq T_0$ and any sequence of queries of T queries the regret is at most $\mathcal{R}_A(T) \leq C \cdot T^\gamma$.*
- *For any $\gamma < \frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+2}$ there exists positive constants C, T_0 such that for any $T \geq T_0$ and any algorithm A there exists an instance μ and a sequence of T queries such that the regret is at least $\mathcal{R}_A(T) \geq C \cdot T^\gamma$.*

If the query space and the ads space are bounded subsets of Euclidean spaces or are finite then $\tilde{a} = a$ and $\tilde{b} = b$ (finite spaces have zero dimension) and the theorem provides matching upper and lower bounds.

Discussion: In this paper we ignore any computational issues. We present the query-ad-clustering algorithm merely as a function or a decision rule, not as a computational procedure. Nevertheless the algorithm can be turned into an efficiently computable procedure with running time polynomial in T provided that one can effectively construct coverings of the query and ads spaces.

Organization of the paper: The paper is organized as follows. In section 1.1 we discuss related work. In section 1.2 we formally define the Lipschitz context multi-armed problem, and we introduce the necessary notations and definitions. In the rest of the paper we prove Theorem 1. The first part of Theorem 1 is proven in section 2 where we present the query-ad-clustering algorithm and prove the upper bound on its regret. The second part of Theorem 1—the lower bound on regret of any algorithm—is proven in section 3.

1.1 Related Work

There is a body of relevant literature on context-*free* multi-armed bandit problems: first bounds on the regret for the model with finite action space were obtained in the classic paper by Lai and Robbins [17]; a more detailed exposition can be found in Auer et al. [2]. A version of the model where the payoffs are chosen adversarially in each round was introduced by Auer, et al. [3]. In recent years much work has been done on very large action spaces. Flaxman et al [10] considered a setting where actions form a convex set and in each round a convex payoff function is adversarially chosen. Continuum actions spaces and payoff functions satisfying (variants of) Lipschitz condition were studied in [13, 14, 5]. Most recently, metric action spaces where the payoff function is Lipschitz was considered by Kleinberg et al. [15]. Inspired by their work, we also consider metric spaces for our work.

Our model can be viewed as a direct and strict generalization of the classical multi-armed bandit problem by Lai and Robbins and the bandit problem in continuum and general metric spaces as presented by Agrawall [1] and Kleinberg et al. [15]. These models can be viewed as a special case of our model where the query space is a singleton. Our upper and lower bounds on the regret apply to these models as well. Compared to Kleinberg et al.'s results [15] whose bounds are in terms of a metric dependent max-min-covering dimension, our lower bound might seem contradictory. However, the important difference is the non-uniformity over the payoff function μ. Namely, our bounds do not depend on μ whereas theirs do.

Online learning model with expert advice is a class of models related to multi-armed bandit problems, see the book by Cesa-Bianchi and Lugosi [6]. These can viewed as multi-armed bandit problems with side information, but their structure is different than the structure of our model.

We are aware of two papers that define multi-armed bandit problem with side information: Wang, Kulkarni and Poor [20] and Goldenshluger and Zeevi [11]. However, the models in these paper are very different from ours.

Regret bounds for reinforcement learning has been studied by several authors. See for example, papers by Auer and Orbert [4], Mansour and Evan-Dar [9]. For a general overview of reinforcement learning see the book [19] by Sutton and Barto.

1.2 Notation

Definition 2. *A* Lipschitz context multi-armed bandit problem *(Lipschitz context MAB) is a pair of metric spaces—a metric space of* queries (X, L_X) *of and a metric space of* ads (Y, L_Y). *An* instance *of the problem is a* payoff *function* $\mu : X \times Y \to [0, 1]$ *which is Lipschitz in each coordinate, that is,*

$$\forall x, x' \in X,\ \forall y, y' \in Y \qquad |\mu(x, y) - \mu(x', y')| \leq L_X(x, x') + L_Y(y, y') \ . \tag{1}$$

The above condition can still be meaningful if the metric spaces have diameter greater than unity, however, we steer clear of the issue of learning meaningful metrics. In the above definition, the Lipschitz condition (1) can be equivalently, perhaps more intuitively, written as a pair of Lipschitz conditions, one condition for the query space and one for the ad space:

$$\begin{aligned} &\forall x, x' \in X,\ \forall y \in Y \qquad |\mu(x, y) - \mu(x', y)| \leq L_X(x, x') \ , \\ &\forall x \in X,\ \forall y, y' \in Y \qquad |\mu(x, y) - \mu(x, y')| \leq L_Y(y, y') \ . \end{aligned}$$

An *algorithm* for a Lipschitz context MAB is a sequence $A = \{A_t\}_{t=1}^{\infty}$ of functions $A_t : (X \times Y \times [0, 1])^{t-1} \times X \to Y$ where the function A_t maps a history $(x_1, y_1, \hat{\mu}_1), (x_2, y_2, \hat{\mu}_2), \ldots, (x_{t-1}, y_{t-1}, \hat{\mu}_{t-1})$ and a current query x_t to an ad y_t. The algorithm operates in rounds $t = 1, 2, \ldots$ in an online fashion. In each round t the algorithm first receives a query x_t, then (based on the query and the history) it displays an ad y_t, and finally it receives payoff[2] $\hat{\mu}_t \in [0, 1]$ which is an independent random variable with expectation $\mu(x_t, y_t)$. *Regret* of A after T rounds on a fixed sequence of queries $x_1, x_2, \ldots, x_T$ is defined as

$$\mathcal{R}_A(T) = \sum_{t=1}^{T} \sup_{y'_t \in Y} \mu(x_t, y'_t) - \mathbf{E}\left[\sum_{t=1}^{T} \mu(x_t, y_t)\right]$$

where the expectation is taken over the random choice of the payoff sequence $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_T$ that the algorithm receives.

Our results are upper and lower bounds on the regret. We express those bounds in terms of covering and packing dimensions of the query space and the ad space, respectively. These dimensions are in turn defined in terms of covering and packing numbers. We specify these notions formally in the following definition.

Definition 3. *Let* (Z, L_Z) *be a metric space.* Covering number $\mathcal{N}(Z, L_Z, r)$ *is the smallest number of sets needed to cover* Z *such that in each set of the covering any two points have distance less than* r. *The* covering dimension *of* (Z, L_Z) *is*

$$\texttt{COV}(Z, L_Z) = \inf\left\{d \ :\ \exists c > 0\ \forall r \in (0, 1] \quad \mathcal{N}(Z, L_Z, r) \leq c r^{-d}\right\} \ .$$

[2] In the case of clicks, $\hat{\mu}_t \in \{0, 1\}$ where $\hat{\mu}_t = 1$ indicates that the user has clicked on the ad. Our results, however, are the same regardless of whether the range of $\hat{\mu}_t$ is $\{0, 1\}$ or $[0, 1]$.

A subset $Z_0 \subseteq Z$ is called r-separated *if for all $z, z' \in Z_0$ we have $L_Z(z, z') \geq r$. The* packing number $\mathcal{M}(Z, L_Z, r)$ *is the largest size of a r-separated subset.* Packing dimension *of* (Z, L_Z) *is*[3]

$$\texttt{PACK}(Z, L_Z) = \sup\left\{d \;:\; \exists c > 0\, \forall r \in (0,1] \quad \mathcal{M}(Z, L_Z, r) \geq c r^{-d}\right\} .$$

In the rest of the paper, when a Lipschitz context MAB (X, Y) is understood, we denote by a, b the covering dimensions of X, Y respectively and we denote by $\tilde{a}, \tilde{b}$ the packing dimension of X, Y respectively.

2 Query-Ad-Clustering Algorithm

In this section we present the *query-ad-clustering* algorithm for the Lipschitz context MAB. Strictly speaking, the algorithm represents, in fact, a class of algorithms, one for each MAB (X, Y) and each $\gamma > \frac{a+b+1}{a+b+2}$. First we present the algorithm and then we prove $O(T^\gamma)$ upper bound on its regret.

Before we state the algorithm we define several parameters that depend on (X, Y) and γ and fully specify the algorithm. Let a, b to be the covering dimensions of X, Y respectively. We define a', b' so that $a' > a$, $b' > b$ and $\gamma > \frac{a'+b'+1}{a'+b'+2}$. We also let c, d be constants such that the covering numbers of X, Y respectively are bounded as $\mathcal{N}(X, r) \leq c r^{-a'}$ and $\mathcal{N}(Y, r) \leq d r^{-b'}$. Existence of such constants c, d is guaranteed by the definition of covering dimension.

Algorithm Description: The algorithm works in phases $i = 0, 1, 2, \ldots$ consisting of 2^i rounds each. Consider a particular phase i, at the beginning of the phase, the algorithm partitions the query space X into disjoint sets (clusters) $X_1, X_2, \ldots, X_N$ each of diameter at most r where

$$r = 2^{-\frac{i}{a'+b'+2}} \qquad \text{and} \qquad N = c \cdot 2^{\frac{a' i}{a'+b'+2}} . \tag{2}$$

The existence of such partition $X_1, X_2, \ldots, X_N$ follows from the assumption that the covering dimension of X is a. Similarly, at the beginning of the phase, the algorithm picks a subset $Y_0 \subseteq Y$ of size K such that each $y \in Y$ is within distance r to a point in Y, where

$$K = d \cdot 2^{\frac{b' i}{a'+b'+2}} . \tag{3}$$

The existence of such Y_0 comes from the fact that the covering dimension of Y is b. (In phase i, the algorithm displays only ads from Y_0.)

In each round t of the current phase i, when a query x_t is received, the algorithm determines the cluster X_j of the partition to which x_t belongs. Fix a cluster X_j. For each ad $y \in Y_0$, let $n_t(y)$ be the number of

[3]Despite their names, the main difference between the covering dimension and packing dimension is *not* that one uses covering numbers and the other packing numbers. In fact any of those dimensions remains the same regardless of whether we use $\mathcal{M}(Z, L_Z, r)$ or $\mathcal{N}(Z, L_Z, r)$ in the definition. This follows from the classical result:

$$\forall r > 0 \qquad \mathcal{M}(Z, L_Z, 2r) \leq \mathcal{N}(Z, L_Z, r) \leq \mathcal{M}(Z, L_Z, r) ,$$

which is usually attributed to Kolmogorov and Tihomirov [16]. The main difference between the two dimensions is that one definition uses infimum and the other uses supremum. Thus, in general,

$$\texttt{PACK}(Z, L_Z) \leq \texttt{COV}(Z, L_Z) .$$

The inequality might be strict when $\mathcal{N}(Z, L_Z, r)$ (or equivalently $\mathcal{M}(Z, L_Z, r)$) as a function of $(1/r)$ oscillates between two polynomials (in $(1/r)$) of two different degrees. In this light, perhaps better names for the dimensions would be *upper* and *lower* covering dimension. For finite spaces or bounded subsets of an Euclidean space, however, this strange behavior does not occur, and the covering and the packing dimension coincide.

times that the ad y has been displayed for a query from X_j during the current phase up to round t and let $\overline{\mu_t}(y)$ be the corresponding empirical average payoff of ad y. If $n_t(y) = 0$ we define $\mu_t(y) = 0$. In round t, the algorithm displays ad $y \in Y_0$ that maximizes the *upper confidence index*

$$I_{t-1}(y) = \overline{\mu}_{t-1}(y) + R_{t-1}(y)$$

where $R_t = \sqrt{\frac{4i}{1+n_t(y)}}$ is the *confidence radius*. Note that in round t the quantities $n_{t-1}(y)$, $\overline{\mu}_{t-1}(y)$, $R_{t-1}(y)$ and $I_{t-1}(y)$ are available to the algorithm. If multiple ads achieve the maximum upper confidence index, we break ties arbitrarily. This finishes the description of the algorithm.

We now bound the regret of the query-ad-clustering algorithm. In Lemma 4 we bound the regret for a cluster of queries during one phase. The regret of all clusters during one phase is bounded in Lemma 5. The resulting $O(T^\gamma)$ bound is stated as Lemma 6. In proof of Lemma 4 we make use of Hoeffding's bound, proof of which can be found in the book [8, Chapter 2] or in the original paper by Hoeffding [12].

Hoeffding's Inequality. *Let $X_1, X_2, \ldots, X_n$ be independent bounded random variables such that X_i, $1 \leq i \leq n$, has support $[a_i, b_i]$. Then for the sum $S = X_1 + X_2 + \cdots + X_n$ we have for any $u \geq 0$,*

$$\Pr\left[|S - \mathbf{E}[S]| \geq u\right] \leq 2\exp\left(-\frac{2u^2}{\sum_{i=1}^n (a_i - b_i)^2}\right) .$$

Lemma 4. *Assume that during phase i, up to step T, n queries were received in a cluster X_j. Then, the contribution of these queries to the regret is bounded as*

$$\mathcal{R}_{i,j}(T) = \mathbf{E}\left[\sum_{\substack{2^i \leq t \leq \min(T, 2^{i+1}) \\ x_t \in X_j}} \sup_{y'_t \in Y} \mu(x_t, y_t) - \mu(x_t, y_t)\right] \leq 6rn + K\left(\frac{16i}{r} + 1\right)$$

where r is the diameter defined in (2) and K is the size of the ads space covering defined in (3).

Proof. For $i = 0$ the bound is trivial. Henceforth we assume $i \geq 1$. Fix an arbitrary query point x_0 in X_j. Let the *good event* be that $\overline{\mu}_t(y) \in [\mu(x_0, y) - R_t(y) - r, \mu(x_0, y) + R_t(y) + r]$ for all $y \in Y$ and all t, $2^i \leq t < \min(T, 2^i)$. The complement of the good event is the *bad event*.

We use Hoeffding's inequality to show that with high probability the good event occurs. Consider any $y \in Y_0$ and any t, $2^i \leq t < T$, for which $n_t(y) \geq 1$. By Lipschitz condition

$$|\mathbf{E}[\overline{\mu}_t(y)] - \mu(x_0, y)| \leq r .$$

Therefore by Hoeffding's inequality

$$\begin{aligned}
&\Pr\left[\overline{\mu}_t(y) \notin [\mu(x_0, y) - R_t(y) - r, \mu(x_0, y) + R_t(y) + r]\right] \\
&\quad \leq \Pr\left[|\overline{\mu}_t(y) - \mathbf{E}[\overline{\mu}_t(y)]| > R_t(y)\right] \\
&\quad \leq 2\exp\left(-2n_t(y)(R_t(y))^2\right) \\
&\quad \leq 2e^{-4i} \\
&\quad \leq 4^{-i}
\end{aligned}$$

and the same inequality, $\Pr\left[\overline{\mu}_t(y) \notin [\mu(x_0,y) - R_t(y) - r, \mu(x_0,y) + R_t(y) + r]\right] \leq 4^{-i}$, holds trivially if $n_t(y) = 0$ since $R_t(y) > 1$. We use the union bound over all $y \in Y_0$ and all t, $2^i \leq t < \min(T, 2^{i+1})$ to bound the probability of the bad event:

$$\Pr\left[\text{bad event}\right] \leq 2^i |Y_0| 4^{-i} \leq K 2^{-i} \ . \tag{4}$$

Now suppose that the good event occurs. Let $\widehat{\mathcal{R}}$ be the actual regret,

$$\widehat{\mathcal{R}} = \sum_{\substack{2^i \leq t \leq \min(T,2^{i+1}) \\ x_t \in X_j}} \left(\sup_{y'_t \in Y} \mu(x_t, y'_t) - \mu(x_t, y_t) \right) \ .$$

Since the algorithm during the phase i displays ads only from Y_0, the actual regret $\widehat{\mathcal{R}}$ can be decomposed as a sum $\widehat{\mathcal{R}} = \sum_{y \in Y_0} \widehat{\mathcal{R}}_y$ where $\widehat{\mathcal{R}}_y$ is the contribution to the regret by displaying the ad y, that is,

$$\widehat{\mathcal{R}}_y = \sum_{\substack{2^i \leq t \leq \min(T,2^{i+1}) \\ x_t \in X_j \\ y_t = y}} \left(\sup_{y'_t \in Y} \mu(x_t, y'_t) - \mu(x_t, y) \right)$$

Fix $y \in Y_0$. Pick *any* $\epsilon > 0$. Let y^* be an ϵ-optimal for query x_0, that is, y^* is such that $\mu(x_0, y^*) \geq \sup_{y \in Y} \mu(x_0, y) - \epsilon$. Let y_0^* be the optimal ad in Y_0 for the query x_0, that is, $y_0^* = \operatorname{argmax}_{y \in Y_0} \mu(x_0, y)$. Lipschitz condition guarantees that for any $x_t \in X_j$

$$\begin{aligned} \sup_{y'_t \in Y} \mu(x_t, y'_t) &\leq \sup_{y \in Y} \mu(x_0, y) + r \leq \mu(x_0, y^*) + r + \epsilon \leq \mu(x_0, y_0^*) + 2r + \epsilon \ , \\ \mu(x_t, y) &\geq \mu(x_0, y) - r \ . \end{aligned}$$

Using the two inequalities the bound on $\widehat{\mathcal{R}}_y$ simplifies to

$$\widehat{\mathcal{R}}_y \leq n_T(y) \left[\mu(x_0, y_0^*) + 3r + \epsilon - \mu(x_0, y) \right] \ .$$

Since ϵ can be chosen arbitrarily small, we have

$$\forall y \in Y_0 \qquad \widehat{\mathcal{R}}_y \leq n_T(y) \left[\mu(x_0, y_0^*) - \mu(x_0, y) + 3r \right] \ . \tag{5}$$

We split the set Y_0 into two subsets, good ads Y_{good} and bad ads Y_{bad}. An ad y is good when $\mu(x_0, y^*) - \mu(x_0, y) \leq 3r$ or it was not displayed (during phase i up to round T for a query in X_j), otherwise the ad is bad. It follows from (5) and the definition of a good ad that

$$\forall y \in Y_{good} \qquad \widehat{\mathcal{R}}_y \leq 6 r n_T(y) \ . \tag{6}$$

For bad ads we use inequality (5) and give an upper bound on $n_T(y)$. To upper bound $n_T(y)$ we use the good event property. According to the definition of the upper confidence index, the good event is equivalent to $I_t(y) \in [\mu(x_0, y) - r, \mu(x_0, y) + 2R_t(y) + r]$ for all $y \in Y$ and all rounds t, $2^i \leq i < T$. Therefore, the good event implies that for any ad y when the upper bound, $\mu(x_0, y) + 2R_{t-1}(y) + r$, on $I_{t-1}(y)$ gets below the lower bound, $\mu(x_0, y_0^*) - r$, on $I_{t-1}(y_0^*)$ the algorithm stops displaying the ad y for queries from X_j.

Therefore, in the last round t when the ad y is displayed to a query in X_j, is $n_{t-1}(y) + 1 = n_t(y) = n_T(y)$ and

$$\mu(x_0, y) + 2R_{t-1}(y) + r \geq \mu(x_0, y_0^*) - r \ .$$

Equivalently,

$$2R_{t-1}(y) \geq \mu(x_0, y_0^*) - \mu(x_0, y) - 2r \ .$$

We substitute the definition of $R_{t-1}(y)$ into this inequality and square both sides of the inequality. (Note that both side are positive.) This gives an upper bound on $n_T(y) = n_{t-1}(y) + 1$:

$$n_T(y) = n_{t-1}(y) + 1 \leq \frac{16i}{\left(\mu(x_0, y_0^*) - \mu(x_0, y) - 2r\right)^2} \ .$$

Combining with (5) we have

$$\begin{aligned}
\widehat{\mathcal{R}}_y &\leq n_T(y) \left[\mu(x_0, y_0^*) - \mu(x_0, y) + 3r\right] \\
&\leq n_T(y) \left[\mu(x_0, y_0^*) - \mu(x_0, y) - 2r\right] + 5r n_T(y) \\
&\leq \frac{16i}{\mu(x_0, y^*) - \mu(x_0, y) - 2r} + 5r n_T(y) \ .
\end{aligned}$$

Using the definition of a bad ad we get that

$$\forall y \in Y_{\text{bad}} \qquad \widehat{\mathcal{R}}_y \leq \frac{16i}{r} + 5r n_T(y) \ . \tag{7}$$

Summing over all ads, both bad and good, we have

$$\begin{aligned}
\widehat{\mathcal{R}} &= \sum_{y \in Y_{\text{good}}} \widehat{\mathcal{R}}_y + \sum_{y \in Y_{\text{bad}}} \widehat{\mathcal{R}}_y \\
&\leq \sum_{y \in Y_{\text{good}}} 6r n_T(y) + \sum_{y \in Y_{\text{bad}}} \left(\frac{16i}{r} + 5r n_T(y)\right) \\
&\leq 6rn + |Y_{\text{bad}}| \frac{16i}{r} \qquad \text{(since } n \leq 2^i\text{)} \\
&\leq 6rn + K \frac{16i}{r} \ .
\end{aligned}$$

Finally, we bound the expected regret

$$\begin{aligned}
\mathcal{R}_{i,j}(T) = \mathbf{E}\left[\widehat{\mathcal{R}}\right] &\leq n \Pr[\text{bad event}] + \left(6rn + K\frac{16i}{r}\right) \Pr[\text{good event}] \\
&\leq nK2^{-i} + 6rn + K\frac{16i}{r} \\
&\leq K + 6rn + K\frac{16i}{r} \\
&\leq 6rn + K\left(\frac{16i}{r} + 1\right) \ .
\end{aligned}$$

□

Lemma 5. *Assume n queries were received up to round T during a phase i (in any cluster). The contribution of these queries to the regret is bounded as*

$$\mathcal{R}_i(T) = \mathbf{E}\left[\sum_{2^i \le t \le \min(T,2^{i+1})} \sup_{y'_t \in Y} \mu(x_t, y'_t) - \mu(x_t, y_t)\right] \le 6rn + NK\left(\frac{16i}{r} + 1\right) .$$

where r is the diamter defined in (2), N is the size of the query covering defined in (2) and K is the size of the ads space covering defined in (3).

Proof. Let denote by n_j the number of queries belonging to cluster X_j. Clearly $n = \sum_{j=1}^{N} n_j$. From the preceding lemma we have

$$\mathcal{R}_i(T) = \sum_{j=1}^{N} \mathcal{R}_{i,j}(T) \le \sum_{j=1}^{N}\left(6rn_j + K\left(\frac{16i}{r} + 1\right)\right) \le 6rn + NK\left(\frac{16i}{r} + 1\right) .$$

□

Lemma 6. *For any $T \ge 0$, the regret of the query-ad-clustering algorithm is bounded as*

$$\mathcal{R}_A(T) \le (24 + 64cd\log_2 T + 4cd)T^{\frac{a'+b'+1}{a'+b'+2}} = O\left(T^{\gamma}\right) .$$

The lemma proves the first part of Theorem 1.

Proof. Let k be the last phase, that is, k is such that $2^k \le T < 2^{k+1}$. In other words $k = \lfloor \log_2 T \rfloor$. We sum the regret over all phases $0, 1, \ldots, k$. We use the preceding lemma and recall that in phase i

$$r = 2^{-\frac{i}{a'+b'+2}} \qquad N = c \cdot 2^{\frac{a'i}{a'+b'+2}} \qquad K = d \cdot 2^{\frac{b'i}{a'+b'+2}} \qquad n \le 2^i .$$

We have

$$\begin{aligned}
\mathcal{R}_A(T) &= \sum_{i=0}^{k} \mathcal{R}_i(T) \\
&\le \sum_{i=0}^{k} 6 \cdot 2^{-\frac{i}{a'+b'+2}} \cdot 2^i + 2^{\frac{a'i}{a'+b'+2}} \cdot d \cdot 2^{\frac{b'i}{a'+b'+2}} \cdot \left(\frac{16i}{2^{-\frac{i}{a'+b'+2}}} + 1\right) \\
&\le \sum_{i=0}^{k} 6 \cdot 2^{i\frac{a'+b'+1}{a'+b'+2}} + 16cd \cdot i \cdot 2^{i\frac{a'+b'+1}{a'+b'+2}} + cd2^{i\frac{a'+b'}{a'+b'+2}} \\
&\le (6 + 16cdk + cd)\sum_{i=0}^{k}\left(2^{\frac{a'+b'+1}{a'+b'+2}}\right)^i \\
&\le (6 + 16cdk + cd)\, 4\left(2^{\frac{a'+b'+1}{a'+b'+2}}\right)^k \\
&\le (24 + 64cd\log_2 T + 4cd)\, T^{\frac{a'+b'+1}{a'+b'+2}} \\
&= O\left(T^{\frac{a'+b'+1}{a'+b'+2}}\log T\right) = O(T^{\gamma}) .
\end{aligned}$$

□

3 The Lower Bound

In this section we prove for any $\gamma < \frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+1}$ lower bound $\Omega(T^\gamma)$ on the regret of any algorithm for a context Lipschitz MAB (X, Y) with $\tilde{a} = \texttt{PACK}(X, L_Z)$, $\tilde{b} = \texttt{COV}(Y, L_Y)$. On the highest level, the main idea of the lower bound is a simple averaging argument. We construct several "hard" instances and we show that the average regret of any algorithm on those instances is $\Omega(T^\gamma)$.

Before we construct the instances we define several parameters that depend on (X, Y) and γ. We define a', b' so that $a' \in [0, \tilde{a}]$, $b' \in [0, \tilde{b}]$ and $\gamma = \frac{a'+b'+1}{a'+b'+2}$. Moreover, if $\tilde{a} > 0$ we ensure that $a' \in (0, \tilde{a})$ and likewise if $\tilde{b} > 0$ we ensure $b' \in (0, \tilde{b})$. Let c, d be constants such that for any $r \in (0, 1]$ there exist $2r$-separated subsets of X, Y of sizes at least $cr^{-a'}$ and $dr^{-b'}$ respectively. Existence of such constants is guaranteed by the definition of the packing dimension. We also use positive constants α, β, C, T_0 that can be expressed in terms of a', b', c, d only. We don't give the formulas for these constants; they can be in principle extracted from the proofs.

Hard instances: Let time horizon T be given. The "hard" instances are constructed as follows. Let $r = \alpha \cdot T^{-1/(a'+b'+2)}$ and $X_0 \subseteq X$, $Y_0 \subseteq Y$ be $2r$-separated subsets of sizes at least $c \cdot r^{-a'}$, $d \cdot r^{-b'}$ respectively. We construct $|Y_0|^{|X_0|}$ instances each defined by a function $v : X_0 \to Y_0$. For each $v \in Y_0^{X_0}$ we define an instance $\mu_v : X \times Y \to [0, 1]$ as follows. First we define μ_v on $X_0 \times Y$ as

$$\mu_v(x_0, y) = 1/2 + \max\{0,\ r - L_Y(y, v(x_0))\} \qquad \text{for any } x_0 \in X_0, y \in Y,$$

and then we make into a Lipschitz function on the whole domain $X \times Y$ as follows. For any $x \in X$ let $x_0 \in X_0$ be the closest point to x and define for any $y \in Y$

$$\mu_v(x, y) = 1/2 + \max\{0,\ r - L_Y(y, v(x_0)) - L_X(x, x_0)\}\ .$$

Furthermore, we assume that in each round t the payoff $\hat{\mu}_t$ the algorithm receives lies in $\{0, 1\}$, that is, $\hat{\mu}_t$ is a Bernoulli random variable with parameter $\mu_v(x_t, y_t)$.

Now, we choose a sequence of T queries. The sequence of queries will consists of $|X_0|$ subsequences, one for each $x_0 \in X_0$, concatenated together. For each $x_0 \in X_0$ the corresponding subsequence consists of $M = \left\lfloor \frac{T}{|X_0|} \right\rfloor$ (or $M = \left\lfloor \frac{T}{|X_0|} \right\rfloor + 1$) copies of x_0. In Lemma 7 we lower bound the contribution of each subsequence to the total regret. The proof of Lemma 7 is an adaptation of the proof Theorem 6.11 from [6, Chapter 6] of a lower bound for the finitely-armed bandit problem. In Lemma 8 we sum the contributions together and give the final lower bound.

Lemma 7. *For $x_0 \in X_0$ consider a sequence of M copies of query x_0. Then for $T \geq T_0$ and for any algorithm A the average regret on this sequence of queries is lower bounded as*

$$\mathcal{R}_{x_0} = \frac{1}{|Y_0|^{|X_0|}} \sum_{v \in Y_0^{X_0}} \mathcal{R}_A^v(M) \geq \beta\sqrt{|Y_0|M}\ ,$$

where $\mathcal{R}_A^v(M)$ denotes the regret on instance μ_v.

Proof. Deferred to Appendix A. □

Lemma 8. *For any algorithm A, there exists an $v \in Y_0^{X_0}$, and an instance μ_v and a sequence of $T \geq T_0$ queries on which regret is at least*

$$\mathcal{R}_A(T) \geq C \cdot T^\gamma$$

Proof. We use the preceding lemma and sum the regret over all $x_0 \in X_0$.

$$\begin{aligned}
\sup_{v \in Y_0^{X_0}} \mathcal{R}_A^v(T) &\geq \frac{1}{|Y_0|^{|X_0|}} \sum_{v \in Y_0^{X_0}} \mathcal{R}_A^v(T) \\
&\geq \sum_{x_0 \in X_0} \mathcal{R}_{x_0} \\
&\geq \beta |X_0| \sqrt{MT} \\
&= \beta |X_0| \sqrt{|Y_0| \left\lfloor \frac{T}{|X_0|} \right\rfloor} \\
&\geq \beta |X_0| \sqrt{|Y_0| \left(\frac{T}{|X_0|} - 1 \right)} \\
&= \beta \sqrt{|Y_0||X_0|T - |Y_0||X_0|^2} \\
&= \beta \sqrt{|Y_0||X_0|T} - \beta |X_0| \sqrt{|Y_0|} \\
&\text{(using } \sqrt{x-y} > \sqrt{x} - \sqrt{y} \text{ for any } x > y > 0) \\
&= \beta \sqrt{d r^{-b'} \cdot c r^{-a'} \cdot T} - \beta c r^{-a'} \sqrt{d r^{-b'}} \\
&= \beta \sqrt{d T^{\frac{b'}{a'+b'+2}} \cdot c T^{\frac{a'}{a'+b'+2}} \cdot T} - \beta c T^{\frac{a'}{a'+b'+2}} \sqrt{d T^{\frac{b'}{a'+b'+2}}} \\
&= \beta \sqrt{cd} \cdot T^{\frac{a'+b'+1}{a'+b'+2}} - \beta c \sqrt{d} \cdot T^{\frac{a'+b'/2}{a'+b'+2}} \\
&\geq \frac{1}{2} \beta \sqrt{cd} \cdot T^{\frac{a'+b'+1}{a'+b'+2}} \\
&\text{(by choosing } T_0 > (2c)^{\frac{a'+b'+2}{b'/2+1}}) \\
&= \frac{1}{2} \beta \sqrt{cd} \cdot T^{\gamma}
\end{aligned}$$

Setting $C = \frac{1}{2}\beta\sqrt{cd}$ finishes the proof. □

4 Conclusions

We have introduced a novel formulation of the problem of displaying relevant web search ads in the form of a Lipschitz context multi-armed bandit problem. This model naturally captures an online scenario where search queries (context) arrive over time and relevant ads must be shown (multi-armed bandit problem) for each query. It is a strict generalization of previously studied multi-armed bandit settings where no side information is given in each round. We believe that our model applies to many other real life scenarios where additional information is available that affects the rewards of the actions.

When the query and ad spaces are endowed with a metric for which the reward function is Lipschitz, we prove upper and lower bounds on the regret with respect to the Bayesian optimal. Specifically, the upper bound $O(T^{\frac{a+b+1}{a+b+2}+\epsilon})$ is dependent on the covering dimension of the query (a) and ad spaces (b) and the lower bound $\Omega(T^{\frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+2}-\epsilon})$ is dependent on the packing dimensions of spaces $(\tilde{a}, \tilde{b})$. For bounded Euclidean spaces and finite sets, these dimensions are equal and imply nearly tight bounds on the regret. The lower bound can be strengthened to $\overset{\infty}{\Omega}(T^{\gamma})$ for any $\gamma < \max\left\{\frac{a+\tilde{b}+1}{a+\tilde{b}+2}, \frac{\tilde{a}+b+1}{\tilde{a}+b+2}\right\}$. So, if either $\tilde{a} = a$ or $\tilde{b} = b$, then

we can still prove a lower bound that matches the upper bound. However, the lower bound will hold "only" for infinitely many time horizons T (as opposed to all horizons). It seems that for Lipschitz context MABs where $\tilde{a} \neq a$ and $\tilde{b} \neq b$ one needs to craft a different notion of dimension, which would somehow capture the growths of covering numbers of both the query space and the ads space.

Our paper raises some intriguing extensions. First, we can explore the setting where queries are coming IID from a fixed distribution (known or unknown). We expect the worst distribution to be uniform over the query space and have the same regret as the adversarial setting. However, what if the query distribution was concentrated in several regions of the space? In web search we would expect some topics to be much hotter than others. It would be interesting to develop algorithms that can exploit this structure. As well, we can use a more refined metric multi-armed bandit algorithm such as the zooming algorithm [15] for more benign reward functions. Further, one can modify the results for an adaptive adversary with access to an algorithm's decisions and is able to change the Lipschitz reward function in each round.

Acknowledgements. We would like to thank Bobby Kleinberg and John Langford and for useful discussions.

References

[1] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund., and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[4] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems 19, (NIPS 2007)*, pages 49–56. MIT Press, 2007.

[5] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory, (COLT 2007)*, pages 454–468. Springer, 2007.

[6] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Willey & Sons, 2nd edition edition, 2006.

[8] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[9] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[10] Abraham D. Flaxman, Adam T. Kalai, and H. Branden McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM*

symposium on Discrete algorithms (SODA 2005), pages 385–394. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2005.

[11] Alexander Goldenshluger and Assaf Zeevi. Performance limitations in bandit problems with side observations. manuscript, 2007. Available at: `http://www2.gsb.columbia.edu/faculty/azeevi/PAPERS/low-excIEEE.pdf`.

[12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[13] Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17, (NIPS 2005)*, pages 697–704. MIT Press, 2005.

[14] Robert D. Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, Massachusetts Institute of Technology, June 2005. Available at: `http://www.cs.cornell.edu/~rdk/papers/RDK_Thesis.pdf`.

[15] Robert D. Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium, STOC 2008*, pages 681–690. Association for Computing Machinery, 2008. Extended version available at: `http://arxiv.org/PS_cache/arxiv/pdf/0809/0809.4882v1.pdf`.

[16] Andrey N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.

[17] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[18] John Langford. How do we get weak action dependence for learning with partial observations? Blog post: `http://hunch.net/?p=421`, 2008.

[19] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[20] Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, May 2005. Available at: `http://arxiv.org/PS_cache/cs/pdf/0501/0501063v1.pdf`.

A Proof of Lemma 7

Think of $v(x_0)$ being uniformly randomly chosen from Y_0 and let $\mathbf{E}$ denote the expectation with respect to both the random choice of $v(x_0)$ and the payoffs $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$. Clearly, the Bayes optimal payoff is

$$\mathbf{E}\left[\sum_{t=1}^{M} \sup_{y'_t \in Y} \mu_v(x_0, y'_t)\right] = M\,\mathbf{E}\left[\sup_{y \in Y} \mu_v(x_0, y)\right] = M\,\mathbf{E}\left[\mu_v(x_0, v(x_0))\right] = M(1/2 + r)\ .$$

The non-trivial part is to upper bound the payoff of A. First, we partition the ads space Y by forming a Voronoi diagram with sites in Y_0. That is, we consider the partition $P = \{S_y \ : \ y \in Y_0\}$ where $S_y \subseteq Y$ is the set of ads which are closer to $y \in Y_0$ than to any other $y' \in Y_0$. We break ties arbitrarily, but we ensure

that P is a partition of Y. Note that since Y_0 is $2r$-separated S_y contains an open ball of radius r centered at y. Also note that for any $y' \in S_y$ the highest payoff $\mu_v(x_0, y)$ is achieved at the Voronoi site y regardless of v. For $y \in Y_0$ let n_y be the random variable denoting the number of times the algorithm displays an ad from S_y.

Now, let for $y \in Y_0$ denote by $\mathbf{E}_y$ the conditional expectation $\mathbf{E}[\cdot \mid v(x_0) = y]$. The expected payoff of A can be bounded as

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{M} \mu_v(x_0, y_t)\right] &= \frac{1}{|Y_0|} \sum_{y \in Y_0} \mathbf{E}_y \left[\sum_{t=1}^{M} \mu_v(x_0, y_t)\right] \\
&\leq \frac{1}{|Y_0|} \sum_{y \in Y_0} \mathbf{E}_y \left[\sum_{y' \in Y_0} n_{y'}\right] \\
&= \frac{1}{|Y_0|} \sum_{y \in Y_0} \mathbf{E}_y \left[M/2 + r n_y\right] \\
&= M/2 + \frac{r}{|Y_0|} \sum_{y \in Y_0} \mathbf{E}_y \, n_y
\end{aligned}$$

Hence,

$$\mathcal{R}_{x_0} \geq r \left(M - \frac{1}{|Y_0|} \sum_{y \in Y_0} \mathbf{E}_y \, n_y \right) \tag{8}$$

and the proof reduces to bounding $\mathbf{E}_y \, n_y$ from above. We do this by comparing the behavior of A on an "completely noisy" independent instance μ' for which $\mu'(x_0, y) = 1/2$ and the payoffs $\hat{\mu}'_1, \hat{\mu}'_2, \ldots, \hat{\mu}'_M$ are i.i.d. Bernoulli random variables with parameter $1/2$ and are independent from $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M, y_1, y_2, \ldots, y_M$ and $v(x_0)$. We denote by $y'_1, y'_2, \ldots, y_M$ the random variables denoting the ads displayed on μ'. For $y \in Y_0$ let n'_y be a random variable denoting the number of times algorithm A displays an ad from S_y for the noisy instance μ'.

For fixed $y \in Y_0$ we define two probability distributions, q and q', over $\{0,1\}^M$ as follows. For any $B = (b_1, b_2, \ldots, b_M) \in \{0,1\}^M$ let

$$q'(B) = 2^{-M} = \Pr[\hat{\mu}'_1 = b_1, \hat{\mu}'_2 = b_2, \ldots, \hat{\mu}'_M = b_M \mid v(x_0) = y]$$

and

$$q(B) = \Pr[\hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_M = b_M \mid v(x_0) = y] \, .$$

Note that the sequence of payoffs received by the algorithm uniquely determines its behavior and hence for any $y \in Y_0$,

$$\mathbf{E}_y[n_y \mid \hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_M = b_M] = \mathbf{E}[n'_y \mid \hat{\mu}'_1 = b_1, \hat{\mu}'_2 = b_2, \ldots, \hat{\mu}'_M = b_M]$$

Consider, for any $y \in Y_0$,

$$\begin{aligned}
\mathbf{E}\, n'_y - \mathbf{E}_y\, n_y &= \sum_{B\in\{0,1\}^M} q(B)\, \mathbf{E}_y[n_y \mid \hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_M = b_M] \\
&\quad - \sum_{B\in\{0,1\}^M} q'(B)\, \mathbf{E}[n'_y \mid \hat{\mu}'_1 = b_1, \hat{\mu}'_2 = b_2, \ldots, \hat{\mu}'_M = b_M] \\
&= \sum_{B\in\{0,1\}^M} (q(B) - q'(B))\, \mathbf{E}_y[n_y \mid \hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_M = b_M] \\
&\le \sum_{\substack{B\in\{0,1\}^M \\ q(B)>q'(B)}} (q(B) - q'(B))\, \mathbf{E}_y[n_y \mid \hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_M = b_M] \\
&\le M \sum_{\substack{B\in\{0,1\}^M \\ q(B)>q'(B)}} (q(B) - q'(B)) \\
&= \frac{M}{2} \sum_{B\in\{0,1\}^M} |q(B) - q'(B)| \qquad (9)
\end{aligned}$$

where the last inequality follows from that $n_y \le M$. The last expression is $M/2$ times the so-called *total variation* (or L_1) distance between the distributions q, q'. It may be bounded by Pinsker's inequality [7, Lemma 11.6.1] which states that

$$\sum_{B\in\{0,1\}^M} |q(B) - q'(B)| \le \sqrt{2D(q'\|q)}\,, \qquad (10)$$

where

$$D(q'\|q) = \sum_{B\in\{0,1\}^m} q'(B) \ln\left(\frac{q'(B)}{q(B)}\right)$$

is the Kullback-Leibler divergence of the distributions q' and q.

We use the chain rule to compute $D(q'\|q)$. First, for a sequence $B = (b_1, b_2, \ldots, b_{t-1}) \in \{0,1\}^{t-1}$, $1 \le t \le M$, and $b \in \{0,1\}$ we denote by

$$q_t(b|B) = \Pr[\hat{\mu}_t = b \mid \hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_{t-1} = b_{t-1}, v(x_0) = y]$$

and

$$q'_t(b|B) = \Pr[\hat{\mu}'_t = b \mid \hat{\mu}'_1 = b_1, \hat{\mu}'_2 = b_2, \ldots, \hat{\mu}'_{t-1} = b_{t-1}, v(x_0) = y]$$

the conditional distributions of t-th payoffs $\hat{\mu}_t$ and $\hat{\mu}'_t$. Note that the event $\hat{\mu}_1 = b_1, \hat{\mu}_2 = b_2, \ldots, \hat{\mu}_{t-1} = b_{t-1}$ on which we are conditioning, is determined by B and in turn this event determines the ad y_t that A displays in t-round on the instances μ_v. We write y_t as $y_t(B)$ to stress this dependence. Hence, by the chain rule

$$\begin{aligned}
D(q'\|q) &= \sum_{t=1}^{M} \frac{1}{2^{t-1}} \sum_{B\in\{0,1\}^{t-1}} D(q'_t(\cdot|B)\|q_t(\cdot|B)) \\
&= \sum_{t=1}^{M} \frac{1}{2^{t-1}} \left(\sum_{\substack{B\in\{0,1\}^{t-1} \\ y_t(B)\in S_{v(x_0)}}} D(q'_t(\cdot|B)\|q_t(\cdot|B)) + \sum_{\substack{B\in\{0,1\}^{t-1} \\ y_t(B)\notin S_y}} D(q'_t(\cdot|B)\|q_t(\cdot|B)) \right)
\end{aligned}$$

where we have split the inner sum into two cases: (i) the ad $y_t(B)$ lies near the "correct" ad y, that is, $y_t(B) \in S_y$ and (ii) the ad y_t does not lie near the "correct" ad, that is, $y_t(B) \notin S_y$.

The second inner sum in the last expression evaluates to zero, since when $y_t(B) \notin S_y$, $q_t(\cdot|B) = q'_t(\cdot|B) = 1/2$ are the same Bernoulli distribution and thus $D(q'_t(\cdot|B)\|q_t(\cdot|B)) = 0$. The terms of the first inner sum can be bounded if we realize that $q_t(\cdot|B)$ is a Bernoulli distribution with parameter $1/2 + s$ where $s = \max\{0, r - L_Y(y_t, y)\} \leq r$ and $q'_t(\cdot|B)$ is a Bernoulli distribution with parameter $1/2$. Hence, for B for which $y_t \in S_y$

$$\begin{aligned} D(q'_t(\cdot|B)\|q_t(\cdot|B)) &= \frac{1}{2}\ln\left(\frac{1/2}{1/2+s}\right) + \frac{1}{2}\ln\left(\frac{1/2}{1/2-s}\right) \\ &= -\frac{1}{2}\ln(1-4s^2) \\ &\leq 8\ln(4/3)s^2 \\ &\leq 8\ln(4/3)r^2\ , \end{aligned}$$

where used the inequality $-\ln(1-x) \leq 4\ln(4/3)x$ for $x \in [0, 1/4]$ which can be proved by checking it for the left and the right end point of the interval and using the convexity of logarithm. We can guarantee that $r \in [0, 1/4]$ by picking T_0 big enough.

$$D(q'\|q) \leq 8\ln(4/3)r^2 \sum_{t=1}^{M} \frac{1}{2^{t-1}} \sum_{B\in\{0,1\}^{t-1}} \mathbf{1}\{y_t(B) \in S_y\} \tag{11}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function.

We combine (9), Pinsker's inequality (10) and the inequality (11) we have just obtained, and we have

$$\begin{aligned} \left(\frac{1}{|Y_0|}\sum_{y\in Y_0} \mathbf{E}_y\, n_y\right) - \frac{M}{|Y_0|} &= \frac{1}{|Y_0|}\sum_{y\in Y_0}\left(\mathbf{E}_y\, n_y - \mathbf{E}\, n'_y\right) \\ &\leq \frac{M}{2}\frac{1}{|Y_0|}\sum_{y\in Y_0}\sqrt{2D(q\|q')} \\ &\leq \frac{M}{2}\frac{1}{|Y_0|}\sum_{y\in Y_0}\sqrt{2\cdot 8\ln(4/3)r^2\sum_{t=1}^{M}\frac{1}{2^{t-1}}\sum_{B\in\{0,1\}^{t-1}}\mathbf{1}\{y_t(B)\in S_y\}} \\ &\leq \frac{M}{2}\sqrt{\frac{1}{|Y_0|}\sum_{y\in Y_0} 2\cdot 8\ln(4/3)r^2\sum_{t=1}^{M}\frac{1}{2^{t-1}}\sum_{B\in\{0,1\}^{t-1}}\mathbf{1}\{y_t(B)\in S_y\}} \\ &\quad \text{(by the inequality between arithmetic and quadratic mean)} \\ &= Mr\sqrt{\frac{4\ln(4/3)}{|Y_0|}\sum_{t=1}^{M}\frac{1}{2^{t-1}}\sum_{B\in\{0,1\}^{t-1}}\sum_{y\in Y_0}\mathbf{1}\{y_t(B)\in S_y\}} \\ &= Mr\sqrt{\frac{4\ln(4/3)}{|Y_0|M}} \end{aligned}$$

where the last equality follows since $\sum_{B\in\{0,1\}^{t-1}}\sum_{y\in Y_0}\mathbf{1}\{y_t(B)\in S_y\}=2^{t-1}$. Therefore, combining with (8) we have

$$\mathcal{R}_{x_0}\geq r\left(M\left(1-\frac{1}{|Y_0|}\right)-M^{3/2}r\sqrt{\frac{4\ln(4/3)}{|Y_0|}}\right)\ .$$

It can be easily verified that $r=\alpha C\sqrt{|Y_0|/M}$ for some constant C lying in the interval $I=[1/(2\sqrt{cd}),2/\sqrt{cd}]$ provided T_0 is big enough. Substituting that for r leads to

$$\mathcal{R}_{x_0}\geq\left(\left(1-\frac{1}{|Y_0|}\right)C\alpha-C^2\alpha^2\sqrt{4\ln(4/3)}\right)\sqrt{M|Y_0|}\ .$$

If $\alpha>0$ is chosen small enough, $|Y_0|\geq 2$ and $\beta=\min_{C\in I}\left(1-\frac{1}{|Y_0|}\right)C\alpha-C^2\alpha^2\sqrt{4\ln(4/3)}$ is positive. This finishes the proof.