Contextual Multi-Armed Bandits

Tyler Lu
tl@cs.toronto.edu
Department of Computer Science
University of Toronto
10 King's College Road,
M5S 3G4 Toronto, ON, Canada

Dávid Pál
dpal@cs.ualberta.ca
Department of Computing Science
University of Alberta
T6G 2E8 Edmonton, AB, Canada

Martin Pál
mpal@google.com
Google, Inc.
76 9th Avenue, 4th Floor
New York, NY 10011, USA

Abstract

We study *contextual* 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 contextual multi-armed bandit problem models a situation where, in a sequence of independent trials, an online algorithm chooses, based on a given *context* (side information), an action 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 so as to maximize the click-through rate (CTR) of the ads displayed. We cast this problem as a contextual 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 convex bounded subsets of Euclidean spaces, this gives an almost matching upper and lower bound.

1 INTRODUCTION

Internet search engines, such as Google, Yahoo! and Microsoft's Bing, 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 (or CTR) of x and y.

We want 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 $\mu(x, y)$ satisfies a Lipschitz condition with respect to each coordinate. Informally, we assume that the CTRs 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 sequence of queries is 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 CTR. Of course, in order to execute it the CTRs must be known. Instead we are interested in algorithms that do not depend on the knowledge of the CTRs and whose performance is still asymptotically the same as that of the Bayes

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

[1] For simplicity we assume that one ad is displayed per query.

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 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.

The main contributions of this paper are 1) a formal model of the Lipschitz contextual bandit problem on metric spaces, 2) a novel, conceptually simple and clean algorithm, which we call *query-ad-clustering*, and 3) lower bounds that show the algorithm is essentially optimal with respect to regret. In particular, the following theorem states our results in our contextual bandit model. Note that 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 $O(r^{-d})$. 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 $\Omega(r^{-\tilde{d}})$.

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

- *For any $\gamma > \frac{a+b+1}{a+b+2}$, the query-ad-clustering algorithm A has the property that there exists constants T_0, C such that for any instance μ, $T \geq T_0$ and sequence of T queries the regret $\mathcal{R}_A(T) \leq C \cdot T^\gamma$.*
- *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 $\mathcal{R}_A(T) \geq C \cdot T^\gamma$.*

If the query space and the ads space are convex 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.

The paper is organized as follows. In section 1.1 we discuss related work, and introduce our Lipschitz contextual multi-armed bandit model in section 1.2. Then we introduce the query-ad-clustering algorithm in section 2 and give an upper bound on its regret. In section 3 we present what is essentially a matching lower bound on the regret of any Lipschitz contextual bandit algorithm, showing that our algorithm is essentially optimal.

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 [1985]; a more detailed exposition can be found in Auer et al. [2002]. Auer et al. [2003] introduced and provided regret optimal algorithms in the non-stochastic bandit problem when payoffs are adversarial. In recent years much work has been done on very large action spaces. Flaxman et al. [2005] 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 Kleinberg [2005a,b], Auer et al. [2007]. Most recently, metric action spaces where the payoff function is Lipschitz was considered by Kleinberg et al. [2008]. Inspired by their work, we also consider metric spaces for our work. In a follow-up paper by Bubeck et al. [2008] the results of Kleinberg et al. [2008] are extended to more general settings.

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 Agrawal [1995] and Kleinberg et al. [2008]. 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. See section 1.3 for a closer comparison with the model of Kleinberg et al. [2008].

Online learning with expert advice is a class of problems related to multi-armed bandits, see the book by Cesa-Bianchi and Lugosi [2006]. These can viewed as multi-armed bandit problems with side information, but their structure is different than the structure of our model. The most relevant work is the Exp4 algorithm of Auer et al. [2003] where experts are simply any multi-armed bandit algorithm, and the goal is to compete against the best expert. In fact this setting and the Exp4 algorithm can be reformulated in our model, which is discussed further at the end of section 2.

We are aware of three papers that define multi-armed bandit problem with side information. The first two are by Wang et al. [2005] and Goldenshluger and Zeevi [2007], however, the models in these papers are very different from ours. The epoch-greedy algorithm proposed in Langford and Zhang [2007] pertains to a setting where contexts arrive i.i.d. and regret is defined relative to the best context-to-action mapping in some fixed class of such mappings. They upper bound the regret of epoch-greedy in terms of an exploitation parameter that makes it hard to compare with our bounds.

Regret bounds for reinforcement learning has been studied by several authors. See, for example, Auer and Ortner [2007], Even-Dar et al. [2006]. For a general overview of reinforcement learning see Sutton and Barto [1998].

1.2 NOTATION

Definition 2. *A* Lipschitz contextual multi-armed bandit problem *(Lipschitz contextual 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,* $\forall x, x' \in X,\ \forall y, y' \in Y,$

$$|\mu(x,y) - \mu(x',y')| \leq L_X(x,x') + L_Y(y,y'). \qquad (1)$$

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:

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

An *algorithm* for a Lipschitz contextual 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

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

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), *denoted* $\texttt{COV}(Z, L_Z)$, *is*

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

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), *denoted* $\texttt{PACK}(Z, L_Z)$, *is*

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

In the rest of the paper, when a Lipschitz contextual 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.

1.3 COMPARISON WITH Kleinberg et al. [2008]

Compared to the results of Kleinberg et al. [2008] whose bounds are in terms of a metric dependent max-min-covering dimension, our lower bound might seem contradictory since our bound also applies to a query space consisting of a singleton. However, the important difference is the non-uniformity over the payoff function μ. Namely, our bounds do not depend on μ whereas theirs do.

For a fixed metric space (Y, L_Y), let $\boldsymbol{\mu}$ be the set of all Lipschitz payoff functions, for any algorithm A, the *regret dimension* as defined by Kleinberg et al. [2008] is

$$\sup_{\boldsymbol{\mu}} \inf_{d \geq 0} \left\{ \exists T_0,\ \forall T > T_0,\ \mathcal{R}_A(T) \leq T^{\frac{d+1}{d+2}} \right\}.$$

It is shown that there exists algorithms that achieve any regret dimension strictly greater than the max-min-covering dimension and no algorithms exist with regret dimension strictly smaller. The infimum and T_0 in the definition of regret dimension "swallows up" constants that can depend on the payoff in $\boldsymbol{\mu}$.

On the other hand, the constants in our regret bound do not depend on the payoff functions. For example, the lower bound says that there exists constants T_0 and C, for all $T > T_0$, any algorithm A satisfies $\mathcal{R}_A(T) \geq C \cdot T^{\frac{\tilde{b}+1}{\tilde{b}+2}}$ when the query space is a singleton and $\tilde{b} = \texttt{PACK}(Y, L_Y)$.

2 QUERY-AD-CLUSTERING ALGORITHM

In this section we present the *query-ad-clustering* algorithm for the Lipschitz contextual 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

[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 $[0,1]$.

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 $\mathcal{N}(X, r) \leq cr^{-a'}$ and $\mathcal{N}(Y, r) \leq dr^{-b'}$. 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

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

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_0, where

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

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 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 [Devroye and Lugosi, 2001, Chapter 2] or in the original paper by Hoeffding [1963].

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*

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

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, $2^i \leq t \leq \min(T, 2^{i+1} - 1)$. The complement of the good event is the *bad event*.

We use Hoeffding's inequality to show that with probability at most $K2^{-i}$ the bad event occurs conditioned on the values of $n_t(y)$ for all $y \in Y_0$ and all t. Since the $K2^{-i}$ bound does not depend on the values of $n_t(y)$, the bad event occurs with at most this probability unconditionally. 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

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

and the same inequality, $\Pr[\overline{\mu}_t(y) \notin [\mu(x_0, y) - R_t(y) - r, \mu(x_0, y) + R_t(y) + r]] \leq 4^{-i}$, 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, $2^i \leq t \leq \min(T, 2^{i+1} - 1)$ to bound the probability of the bad event:

$$\Pr\left[\text{bad event}\right] \leq 2^i |Y_0| 4^{-i} \leq K2^{-i}. \qquad (4)$$

Recall, that we first conditioned on the values $n_t(y)$ and

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

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

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,

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

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 $\mu(x_0, y^*) \ge \sup_{y \in Y} \mu(x_0, y) - \epsilon$. 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 $x_t \in X_j$

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

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

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

Since ϵ can be chosen arbitrarily small, we have

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

We split the set Y_0 into two subsets, good ads Y_{good} and bad ads Y_{bad}. An ad y is good when $\mu(x_0, y^*) - \mu(x_0, y) \le 3r$ 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

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

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, $2^i \le i < T$. 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

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

Equivalently,

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

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 $n_T(y) = n_{t-1}(y) + 1$:

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

Combining with (5) we have

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

Using the definition of a bad ad we get that

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

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

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

Finally, we bound the expected regret

$$\begin{aligned} \mathcal{R}_{i,j}(T) &= \mathbf{E} \left[\widehat{\mathcal{R}} \right] \\ &\le n \Pr[\text{bad event}] + \left(6rn + K \frac{16i}{r} \right) \Pr[\text{good event}] \\ &\le n K 2^{-i} + 6rn + K \frac{16i}{r} \\ &\le K + 6rn + K \frac{16i}{r} \le 6rn + K \left(\frac{16i}{r} + 1 \right) . \end{aligned}$$

□

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*

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

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

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

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

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

The lemma proves the first part of Theorem 1.

Proof. Let k be the last phase, that is, k is such that $2^k \leq T < 2^{k+1}$. 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

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

We have

$$\begin{aligned}
\mathcal{R}_A(T) &= \sum_{i=0}^{k} \mathcal{R}_i(T) \leq \sum_{i=0}^{k} 6 \cdot 2^{-\frac{i}{a'+b'+2}} \cdot 2^i \\
&+ 2^{\frac{a'i}{a'+b'+2}} \cdot d \cdot 2^{\frac{b'i}{a'+b'+2}} \cdot \left(\frac{16i}{2^{-\frac{i}{a'+b'+2}}} + 1\right) \\
&\leq \sum_{i=0}^{k} 6 \cdot 2^{i\frac{a'+b'+1}{a'+b'+2}} + 16icd2^{i\frac{a'+b'+1}{a'+b'+2}} + cd2^{i\frac{a'+b'}{a'+b'+2}} \\
&\leq (6 + 16cdk + cd) \sum_{i=0}^{k} \left(2^{\frac{a'+b'+1}{a'+b'+2}}\right)^i \\
&\leq (6 + 16cdk + cd)\, 4 \left(2^{\frac{a'+b'+1}{a'+b'+2}}\right)^k \\
&\leq (24 + 64cd\log_2 T + 4cd)\, T^{\frac{a'+b'+1}{a'+b'+2}} \\
&= O\left(T^{\frac{a'+b'+1}{a'+b'+2}} \log T\right) = O(T^{\gamma}). \qquad \square
\end{aligned}$$

While the query-ad-clustering algorithm achieves what turns out to be the optimal regret bound, we note that a modification of the Exp4 "experts" algorithm Auer et al. [2003] achieves the same bound (but we discuss the problems with this algorithm below). Each expert is defined by a mapping $f : \{X_1, \ldots, X_N\} \to Y_0$ where given a $x \in X$ finds the appropriate cluster X_x and recommends $f(X_x)$. There are $E = (1/\epsilon^b)^{(1/\epsilon^a)}$ such experts (mappings), and one of them is ϵ-close to the Bayes optimal strategy. The regret bound Auer et al. [2003] for Exp4 gives us $O(\sqrt{TE \log E})$ to the best expert, which has regret ϵT to the Bayes optimal strategy, setting $\epsilon = T^{-1/(a+b+2)}$ we retrieve the same regret upper bound as query-ad-clustering. However, the problem with this algorithm is that it must keep track of an extremely large number, E, of experts while ignoring the structure of our model—it does not exploit the fact that a bandit algorithm can be run for each context "piece" as opposed to each expert.

3 A 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 contextual Lipschitz MAB (X, Y) with $\tilde{a} = \mathtt{PACK}(X, L_Z)$, $\tilde{b} = \mathtt{PACK}(Y, L_Y)$. 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 for any $(x_0, y) \in X_0 \times Y$ as

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

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 Cesa-Bianchi and Lugosi [2006, 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 the full version of the paper. □

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 $x_0 \in X_0$.

$$\sup_{v \in Y_0^{X_0}} \mathcal{R}_A^v(T) \geq \frac{1}{|Y_0|^{|X_0|}} \sum_{v \in Y_0^{X_0}} \mathcal{R}_A^v(T)$$

$$\geq \sum_{x_0 \in X_0} \mathcal{R}_{x_0} \geq \beta|X_0|\sqrt{MT}$$

$$= \beta|X_0|\sqrt{|Y_0| \left\lfloor \frac{T}{|X_0|} \right\rfloor} \geq \beta|X_0|\sqrt{|Y_0| \left(\frac{T}{|X_0|} - 1 \right)}$$

$$\geq \beta\sqrt{|Y_0||X_0|T} - \beta|X_0|\sqrt{|Y_0|}$$

(using $\sqrt{x-y} > \sqrt{x} - \sqrt{y}$ for any $x > y > 0$)

$$= \beta\sqrt{dr^{-b'} \cdot cr^{-a'} \cdot T} - \beta cr^{-a'}\sqrt{dr^{-b'}}$$

$$= \beta\sqrt{cd} \cdot T^{\frac{a'+b'+1}{a'+b'+2}} - \beta c\sqrt{d} \cdot T^{\frac{a'+b'/2}{a'+b'+2}}$$

$$\geq \frac{1}{2}\beta\sqrt{cd} \cdot T^{\frac{a'+b'+1}{a'+b'+2}} = \frac{1}{2}\beta\sqrt{cd} \cdot T^{\gamma}$$

(by choosing $T_0 > (2c)^{\frac{a'+b'+2}{b'/2+1}}$)

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 contextual multi-armed bandit problem. This model naturally captures an online scenario where search queries (contexts) 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.

We present a very natural and conceptually simple algorithm known as query-ad-clustering, which roughly speaking, clusters the contexts into similar regions and runs a multi-armed bandit algorithm for each context cluster. When the query and ad spaces are endowed with a metric for which the reward function is Lipschitz, we prove an upper bound on the regret of query-ad-clustering and a lower bound on the regret of any algorithm showing that query-ad-clustering is 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 i.i.d. 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 Kleinberg et al. [2008] 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

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

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Ad-*

vances in Neural Information Processing Systems 19, (NIPS 2007), pages 49–56. MIT Press, 2007.

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

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

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.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in x-armed bandits. In *NIPS*, pages 201–208, 2008.

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

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

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.

Abraham D. Flaxman, Adam T. Kalai, and H. Brendan 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.

Alexander Goldenshluger and Assaf Zeevi. Performance limitations in bandit problems with side observations. manuscript, 2007.

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

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, 2005a.

Robert D. Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, Massachusetts Institute of Technology, June 2005b.

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.

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

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

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2007.

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

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.