Journal of Machine Learning Research 15 (2014) 2533-2568 Submitted 2/13; Revised 1/14; Published 7/14

Contextual Bandits with Similarity Information

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research New York City
641 6th Ave, 7th floor
New York, NY 10011
USA

Editor: Nicolo Cesa-Bianchi

Abstract

In a multi-armed bandit (MAB) problem, an online algorithm makes a sequence of choices. In each round it chooses from a time-invariant set of alternatives and receives the payoff associated with this alternative. While the case of small strategy sets is by now well-understood, a lot of recent work has focused on MAB problems with exponentially or infinitely large strategy sets, where one needs to assume extra structure in order to make the problem tractable. In particular, recent literature considered information on similarity between arms.

We consider similarity information in the setting of *contextual bandits*, a natural extension of the basic MAB problem where before each round an algorithm is given the *context*—a hint about the payoffs in this round. Contextual bandits are directly motivated by placing advertisements on web pages, one of the crucial problems in sponsored search. A particularly simple way to represent similarity information in the contextual bandit setting is via a *similarity distance* between the context-arm pairs which bounds from above the difference between the respective expected payoffs.

Prior work on contextual bandits with similarity uses "uniform" partitions of the similarity space, so that each context-arm pair is approximated by the closest pair in the partition. Algorithms based on "uniform" partitions disregard the structure of the payoffs and the context arrivals, which is potentially wasteful. We present algorithms that are based on *adaptive* partitions, and take advantage of "benign" payoffs and context arrivals without sacrificing the worst-case performance. The central idea is to maintain a finer partition in high-payoff regions of the similarity space and in popular regions of the context space. Our results apply to several other settings, e.g., MAB with constrained temporal change (Slivkins and Upfal, 2008) and sleeping bandits (Kleinberg et al., 2008a).

Keywords: multi-armed bandits, contextual bandits, regret, Lipschitz-continuity, metric space

1. Introduction

In a multi-armed bandit problem (henceforth, "multi-armed bandit" will be abbreviated as MAB), an algorithm is presented with a sequence of trials. In each round, the algorithm chooses one alternative from a set of alternatives (*arms*) based on the past history, and receives the payoff associated with this alternative. The goal is to maximize the total payoff of the chosen arms. The MAB setting has been introduced in Robbins (1952) and studied intensively since then in operations research, economics and computer science. This setting

is a clean model for the exploration-exploitation trade-off, a crucial issue in sequential decision-making under uncertainty.

One standard way to evaluate the performance of a bandit algorithm is *regret*, defined as the difference between the expected payoff of an optimal arm and that of the algorithm. By now the MAB problem with a small finite set of arms is quite well understood, e.g., see Lai and Robbins (1985); Auer et al. (2002b,a). However, if the arms set is exponentially or infinitely large, the problem becomes intractable unless we make further assumptions about the problem instance. Essentially, a bandit algorithm needs to find a needle in a haystack; for each algorithm there are inputs on which it performs as badly as random guessing.

Bandit problems with large sets of arms have been an active area of investigation in the past decade (see Section 2 for a discussion of related literature). A common theme in these works is to assume a certain *structure* on payoff functions. Assumptions of this type are natural in many applications, and often lead to efficient learning algorithms (Kleinberg, 2005). In particular, a line of work started in Agrawal (1995) assumes that some information on similarity between arms is available.

In this paper, we consider similarity information in the setting of *contextual bandits* (Woodroofe, 1979; Auer, 2002; Wang et al., 2005; Pandey et al., 2007; Langford and Zhang, 2007), a natural extension of the basic MAB problem where before each round an algorithm is given the *context*—a hint about the payoffs in this round. Contextual bandits are directly motivated by the problem of placing advertisements on web pages, one of the crucial problems in sponsored search. One can cast it as a bandit problem so that arms correspond to the possible ads, and payoffs correspond to the user clicks. Then the context consists of information about the page, and perhaps the user this page is served to. Furthermore, we assume that similarity information is available on both the context and the arms. Following the work in Agrawal (1995); Kleinberg (2004); Auer et al. (2007); Kleinberg et al. (2008b) on the (non-contextual) bandits, a particularly simple way to represent similarity information in the contextual bandit setting is via a *similarity distance* between the context-arm pairs, which gives an upper bound on the difference between the corresponding payoffs.

1.1 Our Model: Contextual Bandits with Similarity Information

The contextual bandits framework is defined as follows. Let X be the *context set* and Y be the *arms set*, and let $\mathcal{P} \subset X \times Y$ be the set of feasible context-arms pairs. In each round t, the following events happen in succession:

1. a context $x_t \in X$ is revealed to the algorithm,
2. the algorithm chooses an arm $y_t \in Y$ such that $(x_t, y_t) \in \mathcal{P}$,
3. payoff (reward) $\pi_t \in [0, 1]$ is revealed.

The sequence of context arrivals $(x_t)_{t \in \mathbb{N}}$ is fixed before the first round, and does not depend on the subsequent choices of the algorithm. With *stochastic payoffs*, for each pair $(x, y) \in \mathcal{P}$ there is a distribution $\Pi(x, y)$ with expectation $\mu(x, y)$, so that π_t is an independent sample from $\Pi(x_t, y_t)$. With *adversarial payoffs*, this distribution can change from round to round. For simplicity, we present the subsequent definitions for the stochastic setting only, whereas the adversarial setting is fleshed out later in the paper (Section 8).

In general, the goal of a bandit algorithm is to maximize the total payoff $\sum_{t=1}^{T} \pi_t$, where T is the *time horizon*. In the contextual MAB setting, we benchmark the algorithm's performance in terms of the context-specific "best arm". Specifically, the goal is to minimize the *contextual regret*:

$$R(T) \triangleq \sum_{t=1}^{T} \mu(x_t, y_t) - \mu^*(x_t), \quad \text{where} \quad \mu^*(x) \triangleq \sup_{y \in Y:\,(x,y)\in\mathcal{P}} \mu(x,y).$$

The context-specific best arm is a more demanding benchmark than the best arm used in the "standard" (context-free) definition of regret.

The similarity information is given to an algorithm as a metric space $(\mathcal{P}, \mathcal{D})$ which we call the *similarity space*, such that the following Lipschitz condition[1] holds:

$$|\mu(x,y) - \mu(x',y')| \leq \mathcal{D}((x,y),\ (x',y')). \tag{1}$$

Without loss of generality, $\mathcal{D} \leq 1$. The absence of similarity information is modeled as $\mathcal{D} = 1$.

An instructive special case is the *product similarity space* $(\mathcal{P}, \mathcal{D}) = (X \times Y, \mathcal{D})$, where $(X, \mathcal{D}_{\mathrm{X}})$ is a metric space on contexts (*context space*), and $(Y, \mathcal{D}_{\mathrm{Y}})$ is a metric space on arms (*arms space*), and

$$\mathcal{D}((x,y),\ (x',y')) = \min(1,\ \mathcal{D}_{\mathrm{X}}(x,x') + \mathcal{D}_{\mathrm{Y}}(y,y')). \tag{2}$$

1.2 Prior Work: Uniform Partitions

Hazan and Megiddo (2007) consider contextual MAB with similarity information on contexts. They suggest an algorithm that chooses a "uniform" partition S_{X} of the context space and approximates x_t by the closest point in S_{X}, call it x'_t. Specifically, the algorithm creates an instance $\mathcal{A}(x)$ of some bandit algorithm $\mathcal{A}$ for each point $x \in S_{\mathrm{X}}$, and invokes $\mathcal{A}(x'_t)$ in each round t. The granularity of the partition is adjusted to the time horizon, the context space, and the black-box regret guarantee for $\mathcal{A}$. Furthermore, Kleinberg (2004) provides a bandit algorithm $\mathcal{A}$ for the adversarial MAB problem on a metric space that has a similar flavor: pick a "uniform" partition S_{Y} of the arms space, and run a k-arm bandit algorithm such as EXP3 (Auer et al., 2002b) on the points in S_{Y}. Again, the granularity of the partition is adjusted to the time horizon, the arms space, and the black-box regret guarantee for EXP3.

Applying these two ideas to our setting (with the product similarity space) gives a simple algorithm which we call the *uniform algorithm*. Its contextual regret, even for adversarial payoffs, is

$$R(T) \leq O(T^{1-1/(2+d_{\mathrm{X}}+d_{\mathrm{Y}})})(\log T), \tag{3}$$

where d_{X} is the covering dimension of the context space and d_{Y} is that of the arms space.

1. In other words, μ is a Lipschitz-continuous function on $(X, \mathcal{P})$, with Lipschitz constant $K_{\mathrm{Lip}} = 1$. Assuming $K_{\mathrm{Lip}} = 1$ is without loss of generality (as long as K_{Lip} is known to the algorithm), since we can re-define $\mathcal{D} \leftarrow K_{\mathrm{Lip}}\, D$.

1.3 Our Contributions

Using "uniform" partitions disregards the potentially benign structure of expected payoffs and context arrivals. The central topic in this paper is ***adaptive partitions*** of the similarity space which are adjusted to frequently occurring contexts and high-paying arms, so that the algorithms can take advantage of the problem instances in which the expected payoffs or the context arrivals are "benign" ("low-dimensional"), in a sense that we make precise later.

We present two main results, one for stochastic payoffs and one for adversarial payoffs. For stochastic payoffs, we provide an algorithm called *contextual zooming* which "zooms in" on the regions of the context space that correspond to frequently occurring contexts, and the regions of the arms space that correspond to high-paying arms. Unlike the algorithms in prior work, this algorithm considers the context space and the arms space *jointly*—it maintains a partition of the similarity space, rather than one partition for contexts and another for arms. We develop provable guarantees that capture the "benign-ness" of the context arrivals and the expected payoffs. In the worst case, we match the guarantee (3) for the uniform algorithm. We obtain nearly matching lower bounds using the KL-divergence technique from Auer et al. (2002b); Kleinberg (2004). The lower bound is very general as it holds for every given (product) similarity space *and* for every fixed value of the upper bound.

Our stochastic contextual MAB setting, and specifically the contextual zooming algorithm, can be fruitfully applied beyond the ad placement scenario described above and beyond MAB with similarity information per se. First, writing $x_t = t$ one can incorporate "temporal constraints" (across time, for each arm), and combine them with "spatial constraints" (across arms, for each time). The analysis of contextual zooming yields concrete, meaningful bounds this scenario. In particular, we recover one of the main results in Slivkins and Upfal (2008). Second, our setting subsumes the stochastic *sleeping bandits* problem (Kleinberg et al., 2008a), where in each round some arms are "asleep", i.e., not available in this round. Here contexts correspond to subsets of arms that are "awake". Contextual zooming recovers and generalizes the corresponding result in Kleinberg et al. (2008a). Third, following the publication of a preliminary version of this paper, contextual zooming has been applied to bandit learning-to-rank in Slivkins et al. (2013).

For the adversarial setting, we provide an algorithm which maintains an adaptive partition of the context space and thus takes advantage of "benign" context arrivals. We develop provable guarantees that capture this "benign-ness". In the worst case, the contextual regret is bounded in terms of the covering dimension of the context space, matching (3). Our algorithm is in fact a *meta-algorithm*: given an adversarial bandit algorithm `Bandit`, we present a contextual bandit algorithm which calls `Bandit` as a subroutine. Our setup is flexible: depending on what additional constraints are known about the adversarial payoffs, one can plug in a bandit algorithm from the prior work on the corresponding version of adversarial MAB, so that the regret bound for `Bandit` plugs into the overall regret bound.

1.4 Discussion

Adaptive partitions (of the arms space) for context-free MAB with similarity information have been introduced in Kleinberg et al. (2008b); Bubeck et al. (2011a). This paper further

explores the potential of the zooming technique in Kleinberg et al. (2008b). Specifically, contextual zooming extends this technique to adaptive partitions of the entire similarity space, which necessitates a technically different algorithm and a more delicate analysis. We obtain a clean algorithm for contextual MAB with improved (and nearly optimal) bounds. Moreover, this algorithm applies to several other, seemingly unrelated problems and unifies some results from prior work.

One alternative approach is to maintain a partition of the context space, and run a separate instance of the zooming algorithm from Kleinberg et al. (2008b) on each set in this partition. Fleshing out this idea leads to the meta-algorithm that we present for adversarial payoffs (with `Bandit` being the zooming algorithm). This meta-algorithm is parameterized (and constrained) by a specific a priori regret bound for `Bandit`. Unfortunately, any a priori regret bound for zooming algorithm would be a pessimistic one, which negates its main strength—the ability to adapt to "benign" expected payoffs.

1.5 Map of the Paper

Section 2 is related work, and Section 3 is Preliminaries. Contextual zooming is presented in Section 4. Lower bounds are in Section 5. Some applications of contextual zooming are discussed in Section 6. The adversarial setting is treated in Section 8.

2. Related Work

A proper discussion of the literature on bandit problems is beyond the scope of this paper. This paper follows the line of work on regret-minimizing bandits; a reader is encouraged to refer to Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012) for background. A different (Bayesian) perspective on bandit problems can be found in Gittins et al. (2011).

Most relevant to this paper is the work on bandits with large sets of arms, specifically bandits with similarity information (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Pandey et al., 2007; Kocsis and Szepesvari, 2006; Munos and Coquelin, 2007; Kleinberg et al., 2008b; Bubeck et al., 2011a; Kleinberg and Slivkins, 2010; Maillard and Munos, 2010). Another commonly assumed structure is linear or convex payoffs (e.g., Awerbuch and Kleinberg, 2008; Flaxman et al., 2005; Dani et al., 2007; Abernethy et al., 2008; Hazan and Kale, 2009; Bubeck et al., 2012). Linear/convex payoffs is a much stronger assumption than similarity, essentially because it allows to make strong inferences about far-away arms. Other assumptions have been considered (e.g., Banks and Sundaram, 1992; Berry et al., 1997; Wang et al., 2008; Bubeck and Munos, 2010). The distinction between stochastic and adversarial payoffs is orthogonal to the structural assumption (such as Lipschitz-continuity or linearity). Papers on MAB with linear/convex payoffs typically allow adversarial payoffs, whereas papers on MAB with similarity information focus on stochastic payoffs, with notable exceptions of Kleinberg (2004) and Maillard and Munos (2010).[2]

The notion of structured adversarial payoffs in this paper is less restrictive than the one in Maillard and Munos (2010) (which in turn specializes the notion from linear/convex payoffs), in the sense that the Lipschitz condition is assumed on the expected payoffs rather

than on realized payoffs. This is a non-trivial distinction, essentially because our notion generalizes stochastic payoffs whereas the other one does not.

2.1 Contextual MAB

In Auer (2002) and Chu et al. (2011)[2] payoffs are linear in context, which is a feature vector. Woodroofe (1979); Wang et al. (2005) and Rigollet and Zeevi (2010)[2] study contextual MAB with stochastic payoffs, under the name *bandits with covariates*: the context is a random variable correlated with the payoffs; they consider the case of two arms, and make some additional assumptions. Lazaric and Munos (2009)[2] consider an online labeling problem with stochastic inputs and adversarially chosen labels; inputs and hypotheses (mappings from inputs to labels) can be thought of as "contexts" and "arms" respectively. *Bandits with experts advice* (e.g., Auer 2002) is the special case of contextual MAB where the context consists of experts' advice; the advice of a each expert is modeled as a distributions over arms. All these papers are not directly applicable to the present setting.

Experimental work on contextual MAB includes (Pandey et al., 2007) and (Li et al., 2010, 2011).[2]

Lu et al. (2010)[2] consider the setting in this paper for a product similarity space and, essentially, recover the uniform algorithm and a lower bound that matches (3). The same guarantee (3) can also be obtained as follows. The "uniform partition" described above can be used to define "experts" for a bandit-with-expert-advice algorithm such as EXP4 (Auer et al., 2002b): for each set of the partition there is an expert whose advise is simply an arbitrary arm in this set. Then the regret bound for EXP4 yields (3). Instead of EXP4 one could use an algorithm in McMahan and Streeter (2009)[2] which improves over EXP4 if the experts are not "too distinct"; however, it is not clear if it translates into concrete improvements over (3).

If the context x_t is time-invariant, our setting reduces to the Lipschitz MAB problem as defined in Kleinberg et al. (2008b), which in turn reduces to continuum-armed bandits (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007) if the metric space is a real line, and to MAB with stochastic payoffs (Auer et al., 2002a) if the similarity information is absent.

3. Preliminaries

We will use the notation from the Introduction. In particular, x_t will denote the t-th *context arrival*, i.e., the context that arrives in round t, and y_t will denote the arm chosen by the algorithm in that round. We will use $x_{(1..T)}$ to denote the sequence of the first T context arrivals $(x_1, \ldots, x_T)$. The *badness* of a point $(x,y) \in \mathcal{P}$ is defined as $\Delta(x,y) \triangleq \mu^*(x) - \mu(x,y)$. The context-specific best arm is

$$y^*(x) \in \operatorname{argmax}_{y \in Y:\,(x,y)\in\mathcal{P}}\ \mu(x,y), \tag{4}$$

where ties are broken in an arbitrary but fixed way. To ensure that the max in (4) is attained by some $y \in Y$, we will assume that the similarity space $(\mathcal{P}, \mathcal{D})$ is compact.

2. This paper is concurrent and independent work w.r.t. the preliminary publication of this paper on `arxiv.org`.

Metric spaces. Covering dimension and related notions are crucial throughout this paper. Let $\mathcal{P}$ be a set of points in a metric space, and fix $r > 0$. An *r-covering* of $\mathcal{P}$ is a collection of subsets of $\mathcal{P}$, each of diameter strictly less than r, that cover $\mathcal{P}$. The minimal number of subsets in an r-covering is called the *r-covering number*[3] of $\mathcal{P}$ and denoted $N_r(\mathcal{P})$. The *covering dimension* of $\mathcal{P}$ (with multiplier c) is the smallest d such that $N_r(\mathcal{P}) \leq c\, r^{-d}$ for each $r > 0$. In particular, if S is a subset of Euclidean space then its covering dimension is at most the linear dimension of S, but can be (much) smaller.

Covering is closely related to packing. A subset $S \subset \mathcal{P}$ is an *r-packing* of $\mathcal{P}$ if the distance between any two points in S is at least r. The maximal number of points in an r-packing is called the *r-packing number* and denoted $N_r^{\texttt{pack}}(\mathcal{P})$. It is well-known that r-packing numbers are essentially the same as r-covering numbers, namely $N_{2r}(\mathcal{P}) \leq N_r^{\texttt{pack}}(\mathcal{P}) \leq N_r(\mathcal{P})$.

The *doubling constant* $c_{\text{DBL}}(\mathcal{P})$ of $\mathcal{P}$ is the smallest k such that any ball can be covered by k balls of half the radius. The doubling constant (and *doubling dimension* $\log c_{\text{DBL}}$) was introduced in Heinonen (2001) and has been a standard notion in theoretical computer science literature since Gupta et al. (2003). It was used to characterize tractable problem instances for a variety of problems (e.g., see Talwar, 2004; Kleinberg et al., 2009; Cole and Gottlieb, 2006). It is known that $c_{\text{DBL}}(\mathcal{P}) \geq c\, 2^d$ if d is the covering dimension of $\mathcal{P}$ with multiplier c, and that $c_{\text{DBL}}(\mathcal{P}) \leq 2^d$ if $\mathcal{P}$ is a bounded subset of d-dimensional Euclidean space. A useful observation is that if distance between any two points in S is $> r$, then any ball of radius r contains at most c_{DBL} points of S.

A ball with center x and radius r is denoted $B(x, r)$. Formally, we will treat a ball as a (center, radius) pair rather than a set of points. A function $f : \mathcal{P} \to \mathbb{R}$ if a Lipschitz function on a metric space $(\mathcal{P}, \mathcal{D})$, with Lipschitz constant K_{Lip}, if the *Lipschitz condition* holds: $|f(x) - f(x')| \leq K_{\text{Lip}}\, \mathcal{D}(x, x')$ for each $x, x' \in \mathcal{P}$.

Accessing the similarity space. We assume full and computationally unrestricted access to the similarity information. While the issues of efficient representation thereof are important in practice, we believe that a proper treatment of these issues would be specific to the particular application and the particular similarity metric used, and would obscure the present paper. One clean formal way to address this issue is to assume *oracle access*: an algorithm accesses the similarity space via a few specific types of queries, and invokes an "oracle" that answers such queries.

Time horizon. We assume that the time horizon is fixed and known in advance. This assumption is without loss of generality in our setting. This is due to the well-known *doubling trick* which converts a bandit algorithm with a fixed time horizon into one that runs indefinitely and achieves essentially the same regret bound. Suppose for any fixed time horizon T there is an algorithm $\texttt{ALG}_T$ whose regret is at most $R(T)$. The new algorithm proceeds in phases $i = 1, 2, 3, \ldots$ of duration 2^i rounds each, so that in each phase i a fresh instance of $\texttt{ALG}_{2^i}$ is run. This algorithm has regret $O(\log T)R(T)$ for each round T, and $O(R(T))$ in the typical case when $R(T) \geq T^\gamma$ for some constant $\gamma > 0$.

3. The covering number can be defined via radius-r balls rather than diameter-r sets. This alternative definition lacks the appealing "robustness" property: $N_r(\mathcal{P}') \leq N_r(\mathcal{P})$ for any $\mathcal{P}' \subset \mathcal{P}$, but (other than that) is equivalent for this paper.

4. The Contextual Zooming Algorithm

In this section we consider the contextual MAB problem with stochastic payoffs. We present an algorithm for this problem, called *contextual zooming*, which takes advantage of both the "benign" context arrivals and the "benign" expected payoffs. The algorithm adaptively maintains a partition of the similarity space, "zooming in" on both the "popular" regions on the context space and the high-payoff regions of the arms space.

Contextual zooming extends the (context-free) zooming technique in Kleinberg et al. (2008b), which necessitates a somewhat more complicated algorithm. In particular, selection and activation rules are defined differently, there is a new notion of "domains" and the distinction between "pre-index" and "index". The analysis is more delicate, both the high-probability argument in Claim 3 and the subsequent argument that bounds the number of samples from suboptimal arms. Also, the key step of setting up the regret bounds is very different, especially for the improved regret bounds in Section 4.4.

4.1 Provable Guarantees

Let us define the notions that express the performance of contextual zooming. These notions rely on the packing number $N_r(\cdot)$ in the similarity space $(\mathcal{P}, \mathcal{D})$, and the more refined versions thereof that take into account "benign" expected payoffs and "benign" context arrivals.

Our guarantees have the following form, for some integer numbers $\{N_r\}_{r\in(0,1)}$:

$$R(T) \leq C_0 \inf_{r_0\in(0,1)} \left(r_0 T + \sum_{r=2^{-i}:\, i\in\mathbb{N},\, r_0\leq r\leq 1} \tfrac{1}{r}\, N_r \log T \right). \tag{5}$$

Here and thereafter, $C_0 = O(1)$ unless specified otherwise. In the pessimistic version, $N_r = N_r(\mathcal{P})$ is the r-packing number[4] of $\mathcal{P}$. The main contribution is refined bounds in which N_r is smaller.

For every guarantee of the form (5), call it N_r-*type* guarantee, prior work (e.g., Kleinberg 2004; Kleinberg et al. 2008b; Bubeck et al. 2011a) suggests a more tractable *dimension-type* guarantee. This guarantee is in terms of the *covering-type dimension* induced by N_r, defined as follows:[5]

$$d_c \triangleq \inf\{d > 0 :\ N_r \leq c\, r^{-d} \quad \forall r \in (0,1)\}. \tag{6}$$

Using (5) with $r_0 = T^{-1/(d_c+2)}$, we obtain

$$R(T) \leq O(C_0)\, (c\, T^{1-1/(2+d_c)}\, \log T) \qquad (\forall c > 0). \tag{7}$$

For the pessimistic version ($N_r = N_r(\mathcal{P})$), the corresponding covering-type dimension d_c is the covering dimension of the similarity space. The resulting guarantee (7) subsumes the bound (3) from prior work (because the covering dimension of a product similarity space is

4. Then (5) can be simplified to $R(T) \leq \inf_{r\in(0,1)} O\left(rT + \frac{1}{r}\, N_r(\mathcal{P}) \log T\right)$, asl $N_r(\mathcal{P})$ is non-increasing in r.
5. One standard definition of the covering dimension is (6) for $N_r = N_r(\mathcal{P})$ and $c = 1$. Following Kleinberg et al. (2008b), we include an explicit dependence on c in (6) to obtain a more efficient regret bound (which holds for any c).

$d_{\mathrm{X}} + d_{\mathrm{Y}}$), and extends this bound from product similarity spaces (2) to arbitrary similarity spaces.

To account for "benign" expected payoffs, instead of r-packing number of the entire set $\mathcal{P}$ we consider the r-packing number of a subset of $\mathcal{P}$ which only includes points with near-optimal expected payoffs:

$$\mathcal{P}_{\mu,r} \triangleq \{(x,y) \in \mathcal{P} : \mu^*(x) - \mu(x,y) \leq 12\, r\}. \tag{8}$$

We define the *r-zooming number* as $N_r(\mathcal{P}_{\mu,r})$, the r-packing number of $\mathcal{P}_{\mu,r}$. The corresponding covering-type dimension (6) is called the *contextual zooming dimension*.

The r-zooming number can be seen as an optimistic version of $N_r(\mathcal{P})$: while equal to $N_r(\mathcal{P})$ in the worst case, it can be much smaller if the set of near-optimal context-arm pairs is "small" in terms of the packing number. Likewise, the contextual zooming dimension is an optimistic version of the covering dimension.

Theorem 1 *Consider the contextual MAB problem with stochastic payoffs. There is an algorithm (namely, Algorithm 1 described below) whose contextual regret $R(T)$ satisfies (5) with N_r equal to $N_r(\mathcal{P}_{\mu,r})$, the r-zooming number. Consequently, $R(T)$ satisfies the dimension-type guarantee (7), where d_c is the contextual zooming dimension.*

In Theorem 1, the same algorithm enjoys the bound (7) for each $c > 0$. This is a useful trade-off since different values of c may result in drastically different values of the dimension d_c. On the contrary, the "uniform algorithm" from prior work essentially needs to take the c as input.

Further refinements to take into account "benign" context arrivals are deferred to Section 4.4.

4.2 Description of the Algorithm

The algorithm is parameterized by the time horizon T. In each round t, it maintains a finite collection $\mathcal{A}_t$ of balls in $(\mathcal{P}, \mathcal{D})$ (called *active balls*) which collectively cover the similarity space. Adding active balls is called *activating*; balls stay active once they are activated. Initially there is only one active ball which has radius 1 and therefore contains the entire similarity space.

At a high level, each round t proceeds as follows. Context x_t arrives. Then the algorithm selects an active ball B and an arm y_t such that $(x_t, y_t) \in B$, according to the "selection rule". Arm y_t is played. Then one ball may be activated, according to the "activation rule".

In order to state the two rules, we need to put forward several definitions. Fix an active ball B and round t. Let $r(B)$ be the radius of B. The *confidence radius* of B at time t is

$$\texttt{conf}_t(B) \triangleq 4\sqrt{\frac{\log T}{1 + n_t(B)}}, \tag{9}$$

where $n_t(B)$ is the number of times B has been selected by the algorithm before round t. The *domain* of ball B in round t is a subset of B that excludes all balls $B' \in \mathcal{A}_t$ of strictly smaller radius:

$$\texttt{dom}_t(B) \triangleq B \setminus \left(\bigcup_{B' \in \mathcal{A}_t:\, r(B') < r(B)} B' \right). \tag{10}$$

Algorithm 1 Contextual zooming algorithm.

Input: Similarity space $(\mathcal{P}, \mathcal{D})$ of diameter ≤ 1, $\mathcal{P} \subset X \times Y$. Time horizon T.
Data: collection $\mathcal{A}$ of "active balls" in $(\mathcal{P}, \mathcal{D})$; counters $n(B)$, $\texttt{rew}(B)$ for each $B \in \mathcal{A}$.

Init: $B \leftarrow B(p, 1)$; // center $p \in \mathcal{P}$ is arbitrary
    $\mathcal{A} \leftarrow \{B\}$; $n(B) = \texttt{rew}(B) = 0$
Main loop: for each round t // use definitions (9-12)
    Input context x_t.
        // activation rule
    $\texttt{relevant} \leftarrow \{B \in \mathcal{A} : (x_t, y) \in \texttt{dom}\,(B, \mathcal{A}) \text{ for some arm } y\}$.
    $B \leftarrow \text{argmax}_{B \in \texttt{relevant}}\ I_t(B)$. // ball B is selected
    $y \leftarrow$ any arm y such that $(x_t, y) \in \texttt{dom}\,(B, \mathcal{A})$.
    Play arm y, observe payoff π.
    Update counters: $n(B) \leftarrow n(B) + 1$, $\texttt{rew}(B) \leftarrow \texttt{rew}(B) + \pi$.
        // selection rule
    if $\texttt{conf}(B) \leq \texttt{radius}(B)$ **then**
        $B' \leftarrow B((x_t, y),\ \frac{1}{2}\,\texttt{radius}(B))$ // new ball to be activated
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{B'\}$; $n(B') = \texttt{rew}(B') = 0$.

We will also denote (10) as $\texttt{dom}\,(B, \mathcal{A}_t)$. Ball B is called *relevant* in round t if $(x_t, y) \in \texttt{dom}_t(B)$ for some arm y. In each round, the algorithm selects one relevant ball B. This ball is selected according to a numerical score $I_t(B)$ called *index*. (The definition of index is deferred to the end of this subsection.)

Now we are ready to state the two rules, for every given round t.

- ***selection rule.*** Select a relevant ball B with the maximal index (break ties arbitrarily). Select an arbitrary arm y such that $(x_t, y) \in \texttt{dom}_t(B)$.
- ***activation rule.*** Suppose the selection rule selects a relevant ball B such that $\texttt{conf}_t(B) \leq r(B)$ after this round. Then, letting y be the arm selected in this round, a ball with center (x_t, y) and radius $\frac{1}{2}\,r(B)$ is activated. (B is then called the *parent* of this ball.)

See Algorithm 1 for the pseudocode.

It remains to define the index $I_t(B)$. Let $\texttt{rew}_t(B)$ be the total payoff from all rounds up to $t-1$ in which ball B has been selected by the algorithm. Then the average payoff from B is $\nu_t(B) \triangleq \frac{\texttt{rew}_t(B)}{\max(1,\, n_t(B))}$. The *pre-index* of B is defined as the average $\nu_t(B)$ plus an "uncertainty term":

$$I_t^{\text{pre}}(B) \triangleq \nu_t(B) + r(B) + \texttt{conf}_t(B). \tag{11}$$

The "uncertainty term" in (11) reflects both uncertainty due to a location in the metric space, via $r(B)$, and uncertainty due to an insufficient number of samples, via $\texttt{conf}_t(B)$.

The index of B is obtained by taking a minimum over all active balls B':

$$I_t(B) \triangleq r(B) + \min_{B' \in \mathcal{A}_t} \left(I_t^{\text{pre}}(B') + \mathcal{D}(B, B') \right), \tag{12}$$

where $\mathcal{D}(B, B')$ is the distance between the centers of the two balls.

Discussion. The meaning of index and pre-index is as follows. Both are upper confidence bound (*UCB*, for short) for expected rewards in B. Pre-index is a UCB for $\mu(B)$, the expected payoff from the center of B; essentially, it is the best UCB on $\mu(B)$ that can be obtained from the observations of B alone. The min expression in (12) is an improved UCB on $\mu(B)$, refined using observations from all other active balls. Finally, index is, essentially, the best available UCB for the expected reward of any pair $(x, y) \in B$.

Relevant balls are defined through the notion of the "domain" to ensure the following property: in each round when a parent ball is selected, some other ball is activated. This property allows us to "charge" the regret accumulated in each such round to the corresponding activated ball.

Running time. The running time is dominated by determining which active balls are relevant. Formally, we assume an oracle that inputs context x and a finite sequence $(B, B_1, \ldots, B_n)$ of balls in the similarity space, and outputs an arm y such that $(x, y) \in B \setminus \cup_{j=1}^{n} B_j$ if such arm exists, and `null` otherwise. Then each round t can be implemented via n_t oracle calls with $n < n_t$ balls each, where n_t is the current number of active balls. Letting $f(n)$ denote the running time of one oracle call in terms of n, the running time for each round the algorithm is at most $n_T\, f(n_T)$.

While implementation of the oracle and running time $f(\cdot)$ depend on the specific similarity space, we can provide some upper bounds on n_T. First, a crude upper bound is $n_T \leq T$. Second, letting $\mathcal{F}_r$ be the collection of all active balls of radius r, we prove that $|\mathcal{F}_r|$ is at most N_r, the r-zooming number of the problem instance. Third, $|\mathcal{F}_r| \leq c_{\text{DBL}}\, T r^2$, where c_{DBL} is the doubling constant of the similarity space. (This is because each active ball must be played at least r^{-2} times before it becomes a parent ball, and each parent ball can have at most c_{DBL} children.) Putting this together, we obtain $n_T \leq \sum_r \min(c_{\text{DBL}}\, T r^2,\, N_r)$, where the sum is over all $r = 2^{-j}$, $j \in \mathbb{N}$.

4.3 Analysis of the Algorithm: Proof of Theorem 1

We start by observing that the activation rule ensures several important invariants.

Claim 2 *The following invariants are maintained:*

- *(confidence) for all times t and all active balls B,*

$$\texttt{conf}_t(B) \leq r(B) \text{ if and only if } B \text{ is a parent ball.}$$

- *(covering) in each round t, the domains of active balls cover the similarity space.*
- *(separation) for any two active balls of radius r, their centers are at distance $\geq r$.*

Proof The confidence invariant is immediate from the activation rule.

For the covering invariant, note that $\cup_{B \in \mathcal{A}} \texttt{dom}\,(B, \mathcal{A}) = \cup_{B \in \mathcal{A}} B$ for any finite collection $\mathcal{A}$ of balls in the similarity space. (For each $v \in \cup_{B \in \mathcal{A}} B$, consider a smallest radius ball in $\mathcal{A}$ that contains B. Then $v \in \texttt{dom}\,(B, \mathcal{A})$.) The covering invariant then follows since $\mathcal{A}_t$ contains a ball that covers the entire similarity space.

To show the separation invariant, let B and B' be two balls of radius r such that B is activated at time t, with parent B^{par}, and B' is activated before time t. The center of B

is some point $(x_t, y_t) \in \texttt{dom}\,(B^{\text{par}}, \mathcal{A}_t)$. Since $r(B^{\text{par}}) > r(B')$, it follows that $(x_t, y_t) \notin B'$. ■

Throughout the analysis we will use the following notation. For a ball B with center $(x, y) \in \mathcal{P}$, define the expected payoff of B as $\mu(B) \triangleq \mu(x, y)$. Let $B_t^{\texttt{sel}}$ be the active ball selected by the algorithm in round t. Recall that the *badness* of $(x, y) \in \mathcal{P}$ is defined as $\Delta(x, y) \triangleq \mu^*(x) - \mu(x, y)$.

Claim 3 *If ball B is active in round t, then with probability at least $1 - T^{-2}$ we have that*

$$|\nu_t(B) - \mu(B)| \leq r(B) + \texttt{conf}_t(B). \tag{13}$$

Proof Fix ball V with center (x, y). Let S be the set of rounds $s \leq t$ when ball B was selected by the algorithm, and let $n = |S|$ be the number of such rounds. Then $\nu_t(B) = \frac{1}{n} \sum_{s \in S} \pi_s(x_s, y_s)$.

Define $Z_k = \sum (\pi_s(x_s, y_s) - \mu(x_s, y_s))$, where the sum is taken over the k smallest elements $s \in S$. Then $\{Z_{k \wedge n}\}_{k \in \mathbb{N}}$ is a martingale with bounded increments. (Note that n here is a random variable.) So by the Azuma-Hoeffding inequality with probability at least $1 - T^{-3}$ it holds that $\frac{1}{k} |Z_{k \wedge n}| \leq \texttt{conf}_t(B)$, for each $k \leq T$. Taking the Union Bound, it follows that $\frac{1}{n} |Z_n| \leq \texttt{conf}_t(B)$. Note that $|\mu(x_s, y_s) - \mu(B)| \leq r(B)$ for each $s \in S$, so $|\nu_t(B) - \mu(B)| \leq r(B) + \frac{1}{n} |Z_n|$, which completes the proof. ■

Note that (13) implies $I^{\text{pre}}(B) \geq \mu(B)$, so that $I^{\text{pre}}(B)$ is indeed a UCB on $\mu(B)$.

Call a run of the algorithm *clean* if (13) holds for each round. From now on we will focus on a clean run, and argue deterministically using (13). The heart of the analysis is the following lemma.

Lemma 4 *Consider a clean run of the algorithm. Then $\Delta(x_t, y_t) \leq 14\, r(B_t^{\texttt{sel}})$ in each round t.*

Proof Fix round t. By the covering invariant, $(x_t,\, y^*(x_t)) \in B$ for some active ball B. Recall from (12) that $I_t(B) = r(B) + I^{\text{pre}}(B') + \mathcal{D}(B, B')$ for some active ball B'. Therefore

$$\begin{aligned}
I_t(B_t^{\texttt{sel}}) &\geq I_t(B) = I^{\text{pre}}(B') + r(B) + \mathcal{D}(B, B') && \text{(selection rule, defn of index (12))} \\
&\geq \mu(B') + r(B) + \mathcal{D}(B, B') && \text{(“clean run”)} \\
&\geq \mu(B) + r(B) \geq \mu(x_t,\, y^*(x_t)) = \mu^*(x_t). && \text{(Lipschitz property (1), twice)}
\end{aligned} \tag{14}$$

On the other hand, letting B^{par} be the parent of $B_t^{\texttt{sel}}$ and noting that by the activation rule

$$\max(\mathcal{D}(B_t^{\texttt{sel}}, B^{\text{par}}),\ \texttt{conf}_t(B^{\text{par}})) \leq r(B^{\text{par}}), \tag{15}$$

we can upper-bound $I_t(B_t^{\texttt{sel}})$ as follows:

$$\begin{aligned}
I^{\text{pre}}(B^{\text{par}}) &= \nu_t(B^{\text{par}}) + r(B^{\text{par}}) + \texttt{conf}_t(B^{\text{par}}) && \text{(defn of preindex (11))} \\
&\leq \mu(B^{\text{par}}) + 2\,r(B^{\text{par}}) + 2\,\texttt{conf}_t(B^{\text{par}}) && \text{(“clean run”)} \\
&\leq \mu(B^{\text{par}}) + 4\,r(B^{\text{par}}) && \text{(“parenthood” (15))} \\
&\leq \mu(B_t^{\texttt{sel}}) + 5\,r(B^{\text{par}}) && \text{(Lipschitz property (1))} \qquad (16) \\
I_t(B_t^{\texttt{sel}}) &\leq r(B_t^{\texttt{sel}}) + I^{\text{pre}}(B^{\text{par}}) + \mathcal{D}(B_t^{\texttt{sel}}, B^{\text{par}}) && \text{(defn of index (12))} \\
&\leq r(B_t^{\texttt{sel}}) + I^{\text{pre}}(B^{\text{par}}) + r(B^{\text{par}}) && \text{(“parenthood” (15))} \\
&\leq r(B_t^{\texttt{sel}}) + \mu(B_t^{\texttt{sel}}) + 6\,r(B^{\text{par}}) && \text{(by (16))} \\
&\leq \mu(B_t^{\texttt{sel}}) + 13\,r(B_t^{\texttt{sel}}) && (r(B^{\text{par}}) = 2\,r(B_t^{\texttt{sel}})) \\
&\leq \mu(x_t, y_t) + 14\,r(B_t^{\texttt{sel}}) && \text{(Lipschitz property (1)).} \qquad (17)
\end{aligned}$$

Putting the pieces together, $\mu^*(x_t) \leq I_t(B_t^{\texttt{sel}}) \leq \mu(x_t, y_t) + 14\,r(B_t^{\texttt{sel}})$. ■

Corollary 5 *In a clean run, if ball B is activated in round t then $\Delta(x_t, y_t) \leq 10\,r(B)$.*

Proof By the activation rule, $B_t^{\texttt{sel}}$ is the parent of B. Thus by Lemma 4 we immediately have $\Delta(x_t, y_t) \leq 14\,r(B_t^{\texttt{sel}}) = 28\,r(B)$.

To obtain the constant of 10 that is claimed here, we prove a more efficient special case of Lemma 4:

$$\text{if } B_t^{\texttt{sel}} \text{ is a parent ball then } \Delta(x_t, y_t) \leq 5\,r(B_t^{\texttt{sel}}). \qquad (18)$$

To prove (18), we simply replace (17) in the proof of Lemma 4 by similar inequality in terms of $I^{\text{pre}}(B_t^{\texttt{sel}})$ rather than $I^{\text{pre}}(B^{\text{par}})$:

$$\begin{aligned}
I_t(B_t^{\texttt{sel}}) &\leq r(B_t^{\texttt{sel}}) + I^{\text{pre}}(B_t^{\texttt{sel}}) && \text{(defn of index (12))} \\
&= \nu_t(B_t^{\texttt{sel}}) + 2\,r(B_t^{\texttt{sel}}) + \texttt{conf}_t(B_t^{\texttt{sel}}) && \text{(defns of pre-index (11))} \\
&\leq \mu(B_t^{\texttt{sel}}) + 3\,r(B_t^{\texttt{sel}}) + 2\,\texttt{conf}_t(B_t^{\texttt{sel}}) && \text{(“clean run”)} \\
&\leq \mu(x_t, y_t) + 5\,r(B_t^{\texttt{sel}})
\end{aligned}$$

For the last inequality, we use the fact that $\texttt{conf}_t(B_t^{\texttt{sel}}) \leq r(B_t^{\texttt{sel}})$ whenever $B_t^{\texttt{sel}}$ is a parent ball. ■

Now we are ready for the final regret computation. For a given $r = 2^{-i}$, $i \in \mathbb{N}$, let $\mathcal{F}_r$ be the collection of all balls of radius r that have been activated throughout the execution of the algorithm. Note that in each round, if a parent ball is selected then some other ball is activated. Thus, we can partition the rounds among active balls as follows: for each ball $B \in \mathcal{F}_r$, let S_B be the set of rounds which consists of the round when B was activated and all rounds t when B was selected and was not a parent ball.[6] It is easy to see that

6. A given ball B can be selected even after it becomes a parent ball, but in such round some other ball B is activated, so this round is included in $S_{B'}$.

$|S_B| \leq O(r^{-2} \log T)$. Moreover, by Lemma 4 and Corollary 5 we have $\Delta(x_t, y_t) \leq 15\, r$ in each round $t \in S_B$.

If ball $B \in \mathcal{F}_r$ is activated in round t, then Corollary 5 asserts that its center (x_t, y_t) lies in the set $\mathcal{P}_{\mu,r}$, as defined in (8). By the separation invariant, the centers of balls in $\mathcal{F}_r$ are within distance at least r from one another. It follows that $|\mathcal{F}_r| \leq N_r$, where N_r is the r-zooming number.

Fixing some $r_0 \in (0,1)$, note that in each rounds t when a ball of radius $< r_0$ was selected, regret is $\Delta(x_t, y_t) \leq O(r_0)$, so the total regret from all such rounds is at most $O(r_0\, T)$. Therefore, contextual regret can be written as follows:

$$\begin{aligned}
R(T) &= \textstyle\sum_{t=1}^{T} \Delta(x_t, y_t) \\
&= O(r_0\, T) + \textstyle\sum_{r=2^{-i}:\ r_0 \leq r \leq 1}\ \sum_{B \in \mathcal{F}_r} \sum_{t \in S_B} \Delta(x_t, y_t) \\
&\leq O(r_0\, T) + \textstyle\sum_{r=2^{-i}:\ r_0 \leq r \leq 1}\ \sum_{B \in \mathcal{F}_r} |S_B|\, O(r) \\
&\leq O\left(r_0 T \ + \textstyle\sum_{r=2^{-i}:\ r_0 \leq r \leq 1}\ \frac{1}{r}\, N_r \log(T) \right).
\end{aligned}$$

The N_r-type regret guarantee in Theorem 1 follows by taking inf on all $r_0 \in (0,1)$.

4.4 Improved Regret Bounds

Let us provide regret bounds that take into account "benign" context arrivals. The main difficulty here is to develop the corresponding definitions; the analysis then carries over without much modification. The added value is two-fold: first, we establish the intuition that benign context arrivals matter, and then the specific regret bound is used in Section 6.2 to match the result in Slivkins and Upfal (2008).

A crucial step in the proof of Theorem 1 is to bound the number of active radius-r balls by $N_r(\mathcal{P}_{\mu,r})$, which is accomplished by observing that their centers form an r-packing S of $\mathcal{P}_{\mu,r}$. We make this step more efficient, as follows. An active radius-r ball is called *full* if $\texttt{conf}_t(B) \leq r$ for some round t. Note that each active ball is either full or a child of some other ball that is full. The number of children of a given ball is bounded by the doubling constant of the similarity space. Thus, it suffices to consider the number of active radius-r balls that are full, which is at most $N_r(\mathcal{P}_{\mu,r})$, and potentially much smaller.

Consider active radius-r active balls that are full. Their centers form an r-packing S of $\mathcal{P}_{\mu,r}$ with an additional property: each point $p \in S$ is assigned at least $1/r^2$ context arrivals x_t so that $(x_t, y) \in B(p, r)$ for some arm y, and each context arrival is assigned to at most one point in S.[7] A set $S \subset \mathcal{P}$ with this property is called *r-consistent* (with context arrivals). The *adjusted r-packing number* of a set $\mathcal{P}' \subset \mathcal{P}$, denoted $N_r^{\text{adj}}(\mathcal{P}')$, is the maximal size of an r-consistent r-packing of $\mathcal{P}'$. It can be much smaller than the r-packing number of $\mathcal{P}'$ if most context arrivals fall into a small region of the similarity space.

We make one further optimization, tailored to the application in Section 6.2. Informally, we take advantage of context arrivals x_t such that expected payoff $\mu(x_t, y)$ is either optimal or very suboptimal. A point $(x, y) \in \mathcal{P}$ is called an *r-winner* if for each $(x', y') \in B((x, y),\, 2r)$ it holds that $\mu(x', y') = \mu^*(x')$. Let $\mathcal{W}_{\mu,r}$ be the set of all r-winners. It is easy to see that if B is a radius-r ball centered at an r-winner, and B or its child is se-

7. Each point $p \in S$ is assigned all contexts x_t such that the corresponding ball is chosen in round t.

lected in a given round, then this round does not contribute to contextual regret. Therefore, it suffices to consider (r-consistent) r-packings of $\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r}$.

Our final guarantee is in terms of $N^{\text{adj}}(\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r})$, which we term the *adjusted r-zooming number*.

Theorem 6 *Consider the contextual MAB problem with stochastic payoffs. The contextual regret $R(T)$ of the contextual zooming algorithm satisfies (5), where N_r is the adjusted r-zooming number and C_0 is the doubling constant of the similarity space times some absolute constant. Consequently, $R(T)$ satisfies the dimension-type guarantee (7), where d_c is the corresponding covering-type dimension.*

5. Lower Bounds

We match the upper bound in Theorem 1 up to $O(\log T)$ factors. Our lower bound is very general: it applies to an arbitrary product similarity space, and moreover for a given similarity space it matches, up to $O(\log T)$ factors, any fixed value of the upper bound (as explained below).

We construct a distribution $\mathcal{I}$ over problem instances on a given metric space, so that the lower bound is for a problem instance drawn from this distribution. A single problem instance would not suffice to establish a lower bound because a trivial algorithm that picks arm $y^*(x)$ for each context x will achieve regret 0.

The distribution $\mathcal{I}$ satisfies the following two properties: the upper bound in Theorem 1 is uniformly bounded from above by some number R, and any algorithm must incur regret at least $\Omega(R/\log T)$ in expectation over $\mathcal{I}$. Moreover, we constrict such $\mathcal{I}$ for every possible value of the upper bound in Theorem 1 on a given metric space, i.e., not just for problem instances that are "hard" for this metric space.

To formulate our result, let $R^{\text{UB}}_\mu(T)$ denote the upper bound in Theorem 1, i.e., is the right-hand side of (5) where $N_r = N_r(\mathcal{P}_{\mu,r})$ is the r-zooming number. Let $R^{\text{UB}}(T)$ denote the pessimistic version of this bound, namely right-hand side of (5) where $N_r = N_r(\mathcal{P})$ is the packing number of $\mathcal{P}$.

Theorem 7 *Consider the contextual MAB problem with stochastic payoffs, Let $(\mathcal{P}, \mathcal{D})$ be a product similarity space. Fix an arbitrary time horizon T and a positive number $R \leq R^{UB}(T)$. Then there exists a distribution $\mathcal{I}$ over problem instances on $(\mathcal{P}, \mathcal{D})$ with the following two properties:*

(a) $R^{UB}_\mu(T) \leq O(R)$ for each problem instance in $\texttt{support}(\mathcal{I})$.

(b) for any contextual bandit algorithm it holds that $\mathbb{E}_{\mathcal{I}}[R(T)] \geq \Omega(R/\log T)$,

To prove this theorem, we build on the lower-bounding technique from Auer et al. (2002b), and its extension to (context-free) bandits in metric spaces in Kleinberg (2004). In particular, we use the basic *needle-in-the-haystack* example from Auer et al. (2002b), where the "haystack" consists of several arms with expected payoff $\frac{1}{2}$, and the "needle" is an arm whose expected payoff is slightly higher.

5.1 The Lower-Bounding Construction

Our construction is parameterized by two numbers: $r \in (0, \frac{1}{2}]$ and $N \leq N_r(\mathcal{P})$, where $N_r(\mathcal{P})$ is the r-packing number of $\mathcal{P}$. Given these parameters, we construct a collection $\mathcal{I} = \mathcal{I}_{N,r}$ of $\Theta(N)$ problem instances as follows.

Let $N_{\mathrm{X},r}$ be the r-packing number of X in the context space, and let $N_{\mathrm{Y},r}$ be the r-packing number of Y in the arms space. Note that $N_r(\mathcal{P}) = N_{\mathrm{X},r} \times N_{\mathrm{Y},r}$. For simplicity, let us assume that $N = n_{\mathrm{X}}\, n_{\mathrm{Y}}$, where $1 \leq n_X \leq N_{\mathrm{X},r}$ and $2 \leq n_{\mathrm{Y}} \leq N_{\mathrm{Y},r}$.

An *r-net* is the set S of points in a metric space such that any two points in S are at distance $> r$ from each other, and each point in the metric space is within distance $\leq r$ from some point in S. Recall that any r-net on the context space has size at least $N_{\mathrm{X},r}$. Let S_{X} be an arbitrary set of n_{X} points from one such r-net. Similarly, let S_{Y} be an arbitrary set of n_{Y} points from some r-net on the arms space. The sequence $x_{(1..T)}$ of context arrivals is any fixed permutation over the points in S_{X}, repeated indefinitely.

All problem instances in $\mathcal{I}$ have 0-1 payoffs. For each $x \in S_{\mathrm{X}}$ we construct a needle-in-the-haystack example on the set S_{Y}. Namely, we pick one point $y^*(x) \in S_{\mathrm{Y}}$ to be the "needle", and define $\mu(x, y^*(x)) = \frac{1}{2} + \frac{r}{4}$, and $\mu(x, y) = \frac{1}{2} + \frac{r}{8}$ for each $y \in S_{\mathrm{Y}} \setminus \{y^*(x)\}$. We smoothen the expected payoffs so that far from $S_{\mathrm{X}} \times S_{\mathrm{Y}}$ expected payoffs are $\frac{1}{2}$ and the Lipschitz condition (1) holds:

$$\mu(x, y) \triangleq \max_{(x_0, y_0) \in S_{\mathrm{X}} \times S_{\mathrm{Y}}} \max\left(\, \tfrac{1}{2},\ \mu(x_0, y_0) - \mathcal{D}_{\mathrm{X}}(x, x_0) - \mathcal{D}_{\mathrm{Y}}(y, y_0)\ \right). \tag{19}$$

Note that we obtain a distinct problem instance for each function $y^*(\cdot) : S_{\mathrm{X}} \to S_{\mathrm{Y}}$. This completes our construction.

5.2 Analysis

The useful properties of the above construction are summarized in the following lemma:

Lemma 8 *Fix $r \in (0, \frac{1}{2}]$ and $N \leq N_r(\mathcal{P})$. Let $\mathcal{I} = \mathcal{I}_{N,r}$ and $T_0 = N\, r^{-2}$. Then:*

(i) for each problem instance in $\mathcal{I}$ it holds that $R_\mu^{UB}(T_0) \leq O(N/r)(\log T_0)$.

(ii) any contextual bandit algorithm has regret $\mathbb{E}_{\mathcal{I}}[R(T_0)] \geq \Omega(N/r)$ for a problem instance chosen uniformly at random from $\mathcal{I}$.

For the lower bound in Lemma 8, the idea is that in T rounds each context in S_{X} contributes $\Omega(|S_{\mathrm{Y}}|/r)$ to contextual regret, resulting in total contextual regret $\Omega(N/r)$.

Before we proceed to prove Lemma 8, let us use it to derive Theorem 7. Fix an arbitrary time horizon T and a positive number $R \leq R^{\mathrm{UB}}(T)$. Recall that since $N_r(\mathcal{P})$ is non-increasing in r, for some constant $C > 0$ it holds that

$$R^{\mathrm{UB}}(T) = C \times \inf_{r \in (0,1)} \left(rT + \tfrac{1}{r}\, N_r(\mathcal{P}) \log T \right). \tag{20}$$

Claim 9 *Let $r = \frac{R}{2CT(1+\log T)}$. Then $r \leq \frac{1}{2}$ and $Tr^2 \leq N_r(\mathcal{P})$.*

Proof Denote $k(r) = N_r(\mathcal{P})$ and consider function $f(r) \triangleq k(r)/r^2$. This function is non-increasing in r; $f(1) = 1$ and $f(r) \to \infty$ for $r \to 0$. Therefore there exists $r_0 \in (0, 1)$ such that $f(r_0) \leq T \leq f(r_0/2)$. Re-writing this, we obtain

$$k(r_0) \leq T\, r_0^2 \leq 4\, k(r_0/2).$$

It follows that

$$R \leq R^{\text{UB}}(T) \leq C(Tr_0 + \tfrac{1}{r_0}\, k(r_0) \log T) \leq C\, Tr_0(1 + \log T).$$

Thus $r \leq r_0/2$ and finally $T\, r^2 \leq T\, r_0^2/4 \leq k(r_0/2) \leq k(r) = N_r(\mathcal{P})$. ■

So, Lemma 8 with $r \triangleq \frac{R}{2C\, T(1+\log T)}$ and $N \triangleq T\, r^2$. implies Theorem 7.

5.3 Proof of Lemma 8

Claim 10 *Collection $\mathcal{I}$ consists of valid instances of contextual MAB problem with similarity space $(\mathcal{P}, \mathcal{D})$.*

Proof We need to prove that each problem instance in $\mathcal{P}$ satisfies the Lipschitz condition (1). Assume the Lipschitz condition (1) is violated for some points (x, y), $(x', y') \in X \times Y$. For brevity, let $p = (x, y)$, $p' = (x', y')$, and let us write $\mu(p) \triangleq \mu(x, y)$. Then $|\mu(p) - \mu(p')| > \mathcal{D}(p, p')$.

By (19), $\mu(\cdot) \in [\frac{1}{2}, \frac{1}{2} + \frac{r}{4}]$, so $\mathcal{D}(p, p') < \frac{r}{4}$.

Without loss of generality, $\mu(p) > \mu(p')$. In particular, $\mu(p) > \frac{1}{2}$. Therefore there exists $p_0 = (x_0, y_0) \in S_{\text{X}} \times S_{\text{Y}}$ such that $\mathcal{D}(p, p_0) < \frac{r}{4}$. Then $\mathcal{D}(p', p_0) < \frac{r}{2}$ by triangle inequality.

Now, for any other $p_0' \in S_{\text{X}} \times S_{\text{Y}}$ it holds that $\mathcal{D}(p_0, p_0') > r$, and thus by triangle inequality $\mathcal{D}(p, p_0') > \frac{3r}{4}$ and $\mathcal{D}(p', p_0') > \frac{r}{2}$. It follows that (19) can be simplified as follows:

$$\begin{cases} \mu(p) &= \max(\frac{1}{2},\, \mu(p_0) - \mathcal{D}(p, p_0)), \\ \mu(p') &= \max(\frac{1}{2},\, \mu(p_0) - \mathcal{D}(p', p_0)). \end{cases}$$

Therefore

$$\begin{aligned} |\mu(p) - \mu(p')| &= \mu(p) - \mu(p') \\ &= (\mu(p_0) - \mathcal{D}(p, p_0)) - \max(\tfrac{1}{2},\, \mu(p_0) - \mathcal{D}(p', p_0)) \\ &\leq (\mu(p_0) - \mathcal{D}(p, p_0)) - (\mu(p_0) - \mathcal{D}(p', p_0))) \\ &= \mathcal{D}(p', p_0) - \mathcal{D}(p, p_0) \leq \mathcal{D}(p, p'). \end{aligned}$$

So we have obtained a contradiction. ■

Claim 11 *For each instance in $\mathcal{P}$ and $T_0 = N\, r^{-2}$ it holds that $R_\mu^{UB}(T_0) \leq O(N/r)(\log T_0)$.*

Proof Recall that $R_\mu^{\text{UB}}(T_0)$ is the right-hand side of (5) with $N_r = N_r(\mathcal{P}_{\mu,r})$, where $\mathcal{P}_{\mu,r}$ is defined by (8).

Fix $r' > 0$. It is easy to see that

$$\mathcal{P}_{\mu,\, r'} \subset \cup_{p \in S_{\text{X}} \times S_{\text{Y}}}\, B(p, \tfrac{r}{4}).$$

It follows that $N_{r'}(\mathcal{P}_{\mu,r'}) \leq N$ whenever $r' \geq \frac{r}{4}$. Therefore, taking $r_0 = \frac{r}{4}$ in (5), we obtain

$$R_\mu^{\text{UB}}(T_0) \leq O(rT_0 + \tfrac{N}{r} \log T_0) = O(N/r)(\log T_0).$$

■

Claim 12 *Fix a contextual bandit algorithm $\mathcal{A}$. This algorithm has regret $\mathbb{E}_{\mathcal{I}}[R(T_0)] \geq \Omega(N/r)$ for a problem instance chosen uniformly at random from $\mathcal{I}$, where $T_0 = N\, r^{-2}$.*

Proof Let $R(x,T)$ be the contribution of each context $x \in S_{\mathrm{X}}$ to contextual regret:

$$R(x,T) = \sum_{t:\, x_t = x} \mu^*(x) - \mu(x, y_t),$$

where y_t is the arm chosen by the algorithm in round t. Our goal is to show that $R(x,T_0) \geq \Omega(r\, n_{\mathrm{Y}})$.

We will consider each context $x \in S_{\mathrm{X}}$ separately: the rounds when x arrives form an instance I_x of a context-free bandit problem that lasts for $T_0/n_{\mathrm{X}} = n_{\mathrm{Y}}\, r^{-2}$ rounds, where expected payoffs are given by $\mu(x,\cdot)$ as defined in (19). Let $\mathcal{I}_x$ be the family of all such instances I_x.

A uniform distribution over $\mathcal{I}$ can be reformulated as follows: for each $x \in S_{\mathrm{X}}$, pick the "needle" $y^*(x)$ independently and uniformly at random from S_{Y}. This induces a uniform distribution over instances in $\mathcal{I}_x$, for each context $x \in S_{\mathrm{X}}$. Informally, knowing full or partial information about $y^*(x)$ for some x reveals no information whatsoever about $y^*(x')$ for any $x' \neq x$.

Formally, the contextual bandit algorithm $\mathcal{A}$ induces a bandit algorithm $\mathcal{A}_x$ for I_x, for each context $x \in S_{\mathrm{X}}$: the $\mathcal{A}_x$ simulates the problem instance for $\mathcal{A}$ for all contexts $x' \neq x$ (starting from the "needles" $y^*(x')$ chosen independently and uniformly at random from S_{Y}). Then $\mathcal{A}_x$ has expected regret $R_x(T)$ which satisfies $\mathbb{E}[\, R(T)\,] = \mathbb{E}[\, R(x,T)\,]$, where the expectations on both sides are over the randomness in the respective algorithm and the random choice of the problem instance (resp., from $\mathcal{I}_x$ and from $\mathcal{I}$).

Thus, it remains to handle each $\mathcal{I}_x$ separately: i.e., to prove that the expected regret of any bandit algorithm on an instance drawn uniformly at random from $\mathcal{I}_x$ is at least $\Omega(r\, n_{\mathrm{Y}})$. We use the KL-divergence technique that originated in Auer et al. (2002b). If the set of arms were exactly S_{Y}, then the desired lower bound would follow from Auer et al. (2002b) directly. To handle the problem instances in $\mathcal{I}_x$, we use an extension of the technique from Auer et al. (2002b), which is implicit in Kleinberg (2004) and encapsulated as a stand-alone theorem in Kleinberg et al. (2013). We restate this theorem as Theorem 26 in Appendix A.

It is easy to check that the family $\mathcal{I}_x$ of problem instances satisfies the preconditions in Theorem 26. Fix $x \in S_{\mathrm{X}}$. For a given choice of the "needle" $y^* = y^*(x) \in S_{\mathrm{Y}}$, let $\mu(x, y \,|\, y^*)$. be the expected payoff of each arm y, and let $\nu_{y^*}(\cdot) = \mu(x, \cdot \,|\, y^*)$ be the corresponding payoff function for the bandit instance I_x. Then $\{\nu_{y^*}\}$, $y^* \in S_{\mathrm{Y}}$ is an "(ϵ, k)-ensemble" for $\epsilon = \frac{r}{8}$ and $k = |S_{\mathrm{Y}}|$. ■

6. Applications of Contextual Zooming

We describe several applications of contextual zooming: to MAB with slow adversarial change (Section 6.1), to MAB with stochastically evolving payoffs (Section 6.2), and to the "sleeping bandits" problem (Section 6.3). In particular, we recover some of the main results in Slivkins and Upfal (2008) and Kleinberg et al. (2008a). Also, in Section 6.4 we

discuss a recent application of contextual zooming to bandit learning-to-rank, which has been published in Slivkins et al. (2013).

6.1 MAB with Slow Adversarial Change

Consider the (context-free) adversarial MAB problem in which expected payoffs of each arm change over time *gradually*. Specifically, we assume that expected payoff of each arm y changes by at most σ_y in each round, for some a-priori known *volatilities* σ_y. The algorithm's goal here is continuously adapt to the changing environment, rather than converge to the best fixed mapping from contexts to arms. We call this setting the ***drifting MAB problem***.

Formally, our benchmark is a fictitious algorithm which in each round selects an arm that maximizes expected payoff for the current context. The difference in expected payoff between this benchmark and a given algorithm is called *dynamic regret* of this algorithm. It is easy to see that the worst-case dynamic regret of any algorithm cannot be sublinear in time.[8] We are primarily interested in algorithm's long-term performance, as quantified by *average* dynamic regret $\hat{R}(T) \triangleq R(T)/T$. Our goal is to bound the limit $\lim_{T\to\infty} \hat{R}(T)$ in terms of the parameters: the number of arms and the volatilities σ_y. (In general, such upper bound is non-trivial as long as it is smaller than 1, since all payoffs are at most 1.)

We restate this setting as a contextual MAB problem with stochastic payoffs in which the t-th context arrival is simply $x_t = t$. Then $\mu(t, y)$ is the expected payoff of arm y at time t, and dynamic regret coincides with contextual regret specialized to the case $x_t = t$. Each arm y satisfies a "temporal constraint":

$$|\mu(t, y) - \mu(t', y)| \leq \sigma_y \, |t - t'| \tag{21}$$

for some constant σ_y. To set up the corresponding similarity space $(\mathcal{P}, \mathcal{D})$, let $\mathcal{P} = [T] \times Y$, and

$$\mathcal{D}((t, y), (t', y')) = \min(1, \ \sigma_y \, |t - t'| + \mathbf{1}_{\{y \neq y'\}}). \tag{22}$$

Our solution for the drifting MAB problem is the contextual zooming algorithm parameterized by the similarity space $(\mathcal{P}, \mathcal{D})$. To obtain guarantees for the long-term performance, we run contextual zooming with a suitably chosen time horizon T_0, and restart it every T_0 rounds; we call this version *contextual zooming with period* T_0. Periodically restarting the algorithm is a simple way to prevent the change over time from becoming too large; it suffices to obtain strong provable guarantees.

The general provable guarantees are provided by Theorem 1 and Theorem 6. Below we work out some specific, tractable corollaries.

Corollary 13 *Consider the drifting MAB problem with k arms and volatilities $\sigma_y \equiv \sigma$. Contextual zooming with period T_0 has average dynamic regret $\hat{R}(T) = O(k\sigma \log T_0)^{1/3}$, whenever $T \geq T_0 \geq (\frac{k}{\sigma^2})^{1/3} \log \frac{k}{\sigma}$.*

8. For example, consider problem instances with two arms such that the payoff of each arm in each round is either $\frac{1}{2}$ or $\frac{1}{2} + \sigma$ (and can change from round to round). Over this family of problem instances, dynamic regret in T rounds is at least $\frac{1}{2}\sigma T$.

Proof It suffices to upper-bound regret in a single period. Indeed, if $R(T_0) \leq R$ for any problem instance, then $R(T) \leq R\lceil T/T_0 \rceil$ for any $T > T_0$. It follows that $\hat{R}(T) \leq 2\,\hat{R}(T_0)$. Therefore, from here on we can focus on analyzing contextual zooming itself, rather than contextual zooming with a period.

The main step is to derive the regret bound (5) with a specific upper bound on N_r. We will show that

$$\text{dynamic regret } R(\cdot) \text{ satisfies (5) with } N_r \leq k\,\lceil \tfrac{T\sigma}{r} \rceil. \tag{23}$$

Plugging $N_r \leq k\,(1 + \frac{T\sigma}{r})$ into (5) and taking $r_0 = (k\sigma \log T)^{1/3}$ we obtain[9]

$$R(T) \leq O(T)(k\sigma \log T)^{1/3} + O(\tfrac{k^2}{\sigma})^{1/3}(\log T) \qquad \forall T \geq 1.$$

Therefore, for any $T \geq (\frac{k}{\sigma^2})^{1/3} \log \frac{k}{\sigma}$ we have $\hat{R}(T) = O(k\sigma \log T)^{1/3}$.

It remains to prove (23). We use a pessimistic version of Theorem 1: (5) with $N_r = N_r(\mathcal{P})$, the r-packing number of $\mathcal{P}$. Fix $r \in (0, 1]$. For any r-packing S of $\mathcal{P}$ and each arm y, each time interval I of duration $\Delta_r \triangleq r/\sigma$ provides at most one point for S: there exists at most one time $t \in I$ such that $(t, y) \in S$. Since there are at most $\lceil T/\Delta_r \rceil$ such intervals I, it follows that $N_r(\mathcal{P}) \leq k\,\lceil T/\Delta_r \rceil \leq k\,(1 + T\frac{\sigma}{r})$. ■

The restriction $\sigma_y \equiv \sigma$ is non-essential: it is not hard to obtain the same bound with $\sigma = \frac{1}{k}\sum_y \sigma_y$. Modifying the construction in Section 5 (details omitted from this version) one can show that Corollary 13 is optimal up to $O(\log T)$ factors.

Drifting MAB with spatial constraints. The temporal version ($x_t = t$) of our contextual MAB setting with stochastic payoffs subsumes the drifting MAB problem and furthermore allows to combine the temporal constraints (21) described above (for each arm, across time) with "spatial constraints" (for each time, across arms). To the best of our knowledge, such MAB models are quite rare in the literature.[10] A clean example is

$$\mathcal{D}((t, y),\, (t', y')) = \min(1,\; \sigma\,|t - t'| + \mathcal{D}_{\mathrm{Y}}(y, y')), \tag{24}$$

where $(Y, \mathcal{D}_{\mathrm{Y}})$ is the arms space. For this example, we can obtain an analog of Corollary 13, where the regret bound depends on the covering dimension of the arms space $(Y, \mathcal{D}_{\mathrm{Y}})$.

Corollary 14 *Consider the drifting MAB problem with spatial constraints (24), where σ is the volatility. Let d be the covering dimension of the arms space, with multiplier k. Contextual zooming with period T_0 has average dynamic regret $\hat{R}(T) = O(k\,\sigma \log T_0)^{\frac{1}{d+3}}$, whenever $T \geq T_0 \geq k^{\frac{1}{d+3}}\,\sigma^{-\frac{d+2}{d+3}}\,\log \frac{k}{\sigma}$.*

Remark. We obtain Corollary 13 as a special case by setting $d = 0$.

9. This choice of r_0 minimizes the inf expression in (5) up to constant factors by equating the two summands.
10. The only other MAB model with this flavor that we are aware of, found in Hazan and Kale (2009), combines linear payoffs and bounded "total variation" (aggregate temporal change) of the cost functions.

Proof It suffices to bound $\hat{R}(T_0)$ for (non-periodic) contextual zooming. First we bound the r-covering number of the similarity space $(\mathcal{P}, \mathcal{D})$:

$$N_r(\mathcal{P}) = N_r^{\mathrm{X}}(X) \times N_r^{\mathrm{Y}}(Y) \leq \lceil \tfrac{T\sigma}{r} \rceil \, k \, r^{-d},$$

where $N_r^{\mathrm{X}}(\cdot)$ is the r-covering number in the context space, and $N_r^{\mathrm{Y}}(\cdot)$ is that in the arms space. We worked out the former for Corollary 13. Plugging this into (5) and taking $r_0 = (k \, \sigma \log T)^{1/(3+d)}$, we obtain

$$R(T) \leq O(T)(k\sigma \log T)^{\frac{1}{d+3}} + O\left(k^{\frac{2}{d+3}} \, \sigma^{\frac{d+1}{d+3}} \, \log T \right) \qquad \forall T \geq 1.$$

The desired bound on $\hat{R}(T_0)$ follows easily. ■

6.2 Bandits with Stochastically Evolving Payoffs

We consider a special case of drifting MAB problem in which expected payoffs of each arm evolve over time according to a stochastic process with a uniform stationary distribution. We obtain improved regret bounds for contextual zooming, taking advantage of the full power of our analysis in Section 4.

In particular, we address a version in which the stochastic process is a random walk with step $\pm\sigma$. This version has been previously studied in Slivkins and Upfal (2008) under the name "Dynamic MAB". For the main case ($\sigma_i \equiv \sigma$), our regret bound for Dynamic MAB matches that in Slivkins and Upfal (2008).

To improve the flow of the paper, the proofs are deferred to Appendix 7.

Uniform marginals. First we address the general version that we call *drifting MAB with uniform marginals*. Formally, we assume that expected payoffs $\mu(\cdot, y)$ of each arm y evolve over time according to some stochastic process Γ_y that satisfies (21). We assume that the processes Γ_y, $y \in Y$ are mutually independent, and moreover that the marginal distributions $\mu(t, y)$ are uniform on $[0, 1]$, for each time t and each arm y.[11] We are interested in $\mathbb{E}_\Gamma[\hat{R}(T)]$, average dynamic regret in expectation over the processes Γ_y.

We obtain a stronger version of (23) via Theorem 6. To use this theorem, we need to bound the adjusted r-zooming number, call it N_r. We show that

$$\mathbb{E}_\Gamma[N_r] = O(kr)\lceil \tfrac{T\sigma}{r} \rceil \text{ and } \left(r < \sigma^{1/3} \Rightarrow N_r = 0 \right). \tag{25}$$

Then we obtain a different bound on dynamic regret, which is stronger than Corollary 13 for $k < \sigma^{-1/2}$.

Corollary 15 *Consider drifting MAB with uniform marginals, with k arms and volatilities $\sigma_y \equiv \sigma$. Contextual zooming with period T_0 satisfies $\mathbb{E}_\Gamma[\hat{R}(T)] = O(k \, \sigma^{2/3} \, \log T_0)$, whenever $T \geq T_0 \geq \sigma^{-2/3} \log \frac{1}{\sigma}$.*

11. For example, this assumption is satisfied by any Markov Chain on $[0, 1]$ with stationary initial distribution.

The crux of the proof is to show (25). Interestingly, it involves using all three optimizations in Theorem 6: $N_r(\mathcal{P}_{\mu,r})$, $N_r(\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r})$ and $N_r^{\text{adj}}(\cdot)$, whereas any two of them do not seem to suffice. The rest is a straightforward computation similar to the one in Corollary 13.

Dynamic MAB. Let us consider the Dynamic MAB problem from Slivkins and Upfal (2008). Here for each arm y the stochastic process Γ_y is a random walk with step $\pm\sigma_y$. To ensure that the random walk stays within the interval $[0,1]$, we assume reflecting boundaries. Formally, we assume that $1/\sigma_y \in \mathbb{N}$, and once a boundary is reached, the next step is deterministically in the opposite direction.[12]

According to a well-known fact about random walks,[13]

$$\Pr\left[|\mu(t,y) - \mu(t',y)| \leq O(\sigma_y\,|t-t'|^{1/2}\log T_0)\right] \geq 1 - T_0^{-3} \quad \text{if } |t-t'| \leq T_0. \tag{26}$$

We use contextual zooming with period T_0, but we parameterize it by a different similarity space $(\mathcal{P}, \mathcal{D}_{T_0})$ that we define according to (26). Namely, we set

$$\mathcal{D}_{T_0}((t,y),\,(t',y')) = \min(1,\ \sigma_y\,|t-t'|^{1/2}\log T_0 + \mathbf{1}_{\{y\neq y'\}}). \tag{27}$$

The following corollary is proved using the same technique as Corollary 15:

Corollary 16 *Consider the Dynamic MAB problem with k arms and volatilities $\sigma_y \equiv \sigma$. Let $\texttt{ALG}_{T_0}$ denote the contextual zooming algorithm with period T_0 which is parameterized by the similarity space $(\mathcal{P}, \mathcal{D}_{T_0})$. Then $\texttt{ALG}_{T_0}$ satisfies $\mathbb{E}_\Gamma[\hat{R}(T)] = O(k\,\sigma\,\log^2 T_0)$, whenever $T \geq T_0 \geq \frac{1}{\sigma}\log\frac{1}{\sigma}$.*

6.3 Sleeping Bandits

The *sleeping bandits* problem Kleinberg et al. (2008a) is an extension of MAB where in each round some arms can be "asleep", i.e., not available in this round. One of the main results in Kleinberg et al. (2008a) is on sleeping bandits with stochastic payoffs. We recover this result using contextual zooming.

We model sleeping bandits as contextual MAB problem where each context arrival x_t corresponds to the set of arms that are "awake" in this round. More precisely, for every subset $S \subset Y$ of arms there is a distinct context x_S, and $\mathcal{P} = \{(x_S, y) : y \in S \subset Y\}$. is the set of feasible context-arm pairs. The similarity distance is simply $\mathcal{D}((x,y),\,(x',y')) = \mathbf{1}_{\{y\neq y'\}}$. Note that the Lipschitz condition (1) is satisfied.

For this setting, contextual zooming essentially reduces to the "highest awake index" algorithm in Kleinberg et al. (2008a). In fact, we can re-derive the result Kleinberg et al. (2008a) on sleeping MAB with stochastic payoffs as an easy corollary of Theorem 1.

Corollary 17 *Consider the sleeping MAB problem with stochastic payoffs. Order the arms so that their expected payoffs are $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$, where n is the number of arms. Let $\Delta_i = \mu_{i+1} - \mu_i$. Then*

$$R(T) \leq \inf_{r>0}\left(rT + \sum_{i:\,\Delta_i > r} \frac{O(\log T)}{\Delta_i}\right).$$

12. Slivkins and Upfal (2008) has a slightly more general setup which does not require $1/\sigma_y \in \mathbb{N}$.
13. For example, this follows as a simple application of Azuma-Hoeffding inequality.

Proof The r-zooming number $N_r(\mathcal{P}_{\mu,r})$ is equal to the number of distinct *arms* in $\mathcal{P}_{\mu,r}$, i.e., the number of arms $i \in Y$ such that $\Delta(x,i) \leq 12r$ for some context x. Note that for a given arm i, the quantity $\Delta(x,i)$ is minimized when the set of awake arms is $S = \{i, i+1\}$. Therefore, $N_r(\mathcal{P}_{\mu,r})$ is equal to the number of arms $i \in Y$ such that $\Delta_i \leq 12r$. It follows that

$$\begin{aligned}
N_{r>r_0}(\mathcal{P}_{\mu,r}) &= \textstyle\sum_{i=1}^{n} \mathbf{1}_{\{\Delta_i \leq 12r\}}. \\
\textstyle\sum_{r>r_0} \frac{1}{r} N_{r>r_0}(\mathcal{P}_{\mu,r}) &= \textstyle\sum_{r>r_0} \sum_{i=1}^{n} \frac{1}{r} \mathbf{1}_{\{\Delta_i \leq 12r\}} \\
&= \textstyle\sum_{i=1}^{n} \sum_{r>r_0} \frac{1}{r} \mathbf{1}_{\{\Delta_i \leq 12r\}} \\
&= \textstyle\sum_{i:\, \Delta_i > r_0} O(\frac{1}{\Delta_i}). \\
R(T) &\leq \inf_{r_0>0} \left(r_0\, T + O(\log T) \textstyle\sum_{r>r_0} \frac{1}{r} N_r(\mathcal{P}_{\mu,r}) \right) \\
&\leq \inf_{r_0>0} \left(r_0\, T + O(\log T) \textstyle\sum_{i:\, \Delta_i > r_0} O(\frac{1}{\Delta_i}) \right),
\end{aligned}$$

as required. (In the above equations, $\sum_{r>r_0}$ denotes the sum over all $r = 2^{-j} > r_0$ such that $j \in \mathbb{N}$.) ■

Moreover, the contextual MAB problem extends the sleeping bandits setting by incorporating similarity information on arms. The contextual zooming algorithm (and its analysis) applies, and is geared to exploit this additional similarity information.

6.4 Bandit Learning-to-Rank

Following a preliminary publication of this paper on `arxiv.org`, contextual zooming has been applied in Slivkins et al. (2013) to bandit learning-to-rank. Interestingly, the "contexts" studied in Slivkins et al. (2013) are very different from what we considered so far.

The basic setting, motivated by web search, was introduced in Radlinski et al. (2008). In each round a new user arrives. The algorithm selects a ranked list of k documents and presents it to the user who clicks on at most one document, namely on the first document that (s)he finds relevant. A user is specified by a binary vector over documents. The goal is to minimize *abandonment*: the number of rounds with no clicks.

Slivkins et al. (2013) study an extension in which metric similarity information is available. They consider a version with *stochastic payoffs*: in each round, the user vector is an independent sample from a fixed distribution, and assume a Lipschitz-style condition that connects expected clicks with the metric space. They run a separate bandit algorithm (e.g., contextual zooming) for each of the k "slots" in the ranking. Without loss of generality, in each round the documents are selected sequentially, in the top-down order. Since a document in slot i is clicked in a given round only if all higher ranked documents are not relevant, they treat the set of documents in the higher slots as a *context* for the i-th algorithm. The Lipschitz-style condition on expected clicks suffices to guarantee the corresponding Lipschitz-style condition on contexts.

7. Bandits with Stochastically Evolving Payoffs: Missing Proofs

We prove Corollary 15 and Corollary 16 which address the performance of contextual zooming for the stochastically evolving payoffs. In each corollary we bound from above the average dynamic regret $\hat{R}(T)$ of contextual zooming with period T_0, for any $T \geq T_0$. Since $\hat{R}(T) \leq 2\hat{R}(T_0)$, it suffices to bound $\hat{R}(T_0)$, which is the same as $\hat{R}(T_0)$ for (non-periodic) contextual zooming. Therefore, we can focus on analyzing the non-periodic algorithm.

We start with two simple auxiliary claims.

Claim 18 *Consider the contextual MAB problem with a product similarity space. Let $\Delta(x,y) \triangleq \mu^*(x) - \mu(x,y)$ be the "badness" of point (x,y) in the similarity space. Then*

$$|\Delta(x,y) - \Delta(x',y)| \leq 2\,\mathcal{D}_X(x,x') \qquad \forall x,x' \in X,\ y \in Y. \tag{28}$$

Proof First we show that the benchmark payoff $\mu(\cdot)$ satisfies a Lipschitz condition:

$$|\mu^*(x) - \mu^*(x')| \leq \mathcal{D}_{\mathrm{X}}(x,x') \qquad \forall x,x' \in X. \tag{29}$$

Indeed, it holds that $\mu^*(x) = \mu(x,y)$ and $\mu^*(x') = \mu(x,y')$ for some arms $y, y' \in Y$. Then

$$\mu^*(x) = \mu(x,y) \geq \mu(x,y') \geq \mu(x',y') - \mathcal{D}_{\mathrm{X}}(x,x') = \mu^*(x') - \mathcal{D}_{\mathrm{X}}(x,x'),$$

and likewise for the other direction. Now,

$$|\Delta(x,y) - \Delta(x',y)| \leq |\mu^*(x) - \mu^*(x')| + |\mu(x,y) - \mu(x',y)| \leq 2\,\mathcal{D}_{\mathrm{X}}(x,x').$$

■

Claim 19 *Let $Z_1, \ldots, Z_k$ be independent random variables distributed uniformly at random on $[0,1]$. Let $Z^* = \max_i Z_i$. Fix $r > 0$ and let $S = \{i : Z^* > Z_i \geq Z^* - r\}$. Then $\mathbb{E}[\,|S|\,] = kr$.*

This is a textbook result; we provide a proof for the sake of completeness.

Proof Conditional on Z^*, it holds that

$$\begin{aligned}
\mathbb{E}[\,|S|\,] &= \mathbb{E}\left[\textstyle\sum_i \mathbf{1}_{\{Z_i \in S\}}\right] = k\,\Pr[Z_i \in S] \\
&= k\,\Pr[Z_i \in S \,|\, Z_i < Z^*] \times \Pr[Z_i < Z^*] \\
&= k\,\tfrac{r}{Z^*}\,\tfrac{k-1}{k} = (k-1)r/Z^*.
\end{aligned}$$

Integrating over Z^*, and letting $F(z) \triangleq \Pr[Z^* \leq z] = z^k$, we obtain that

$$\begin{aligned}
\mathbb{E}[\,\tfrac{1}{Z^*}\,] &= \textstyle\int_0^1 \frac{1}{z}\,F'(z)dz = \frac{k}{k-1} \\
\mathbb{E}[\,|S|\,] &= (k-1)r\;\mathbb{E}[\,\tfrac{1}{Z^*}\,] = kr.
\end{aligned}$$

■

Proof of Corollary 15 It suffices to bound $\hat{R}(T_0)$ for (non-periodic) contextual zooming.

Let $\mathcal{D}_{\mathrm{X}}(t,t') \triangleq \sigma|t-t'|$ be the context distance implicit in the temporal constraint (21). For each $r>0$, pick a number T_r such that $\mathcal{D}_{\mathrm{X}}(t,t') \leq r \iff |t-t'| \leq T_r$. Clearly, $T_r \triangleq \frac{r}{\sigma}$.

The crux is to bound the adjusted r-zooming number, call it N_r, namely to show (25). For the sake of convenience, let us restate it here (and let us use the notation T_r):

$$\mathbb{E}_\Gamma[N_r] = O(kr)\lceil \tfrac{T}{T_r} \rceil \text{ and } \left(T_r < 1/r^2 \Rightarrow N_r = 0\right). \tag{30}$$

Recall that $N_r = N^{\mathrm{adj}}(\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r})$, where $\mathcal{W}_{\mu,r}$ is the set of all r-winners (see Section 4.4 for the definition). Fix $r \in (0,1]$ and let S be some r-packing of $\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r}$. Partition the time into $\lceil \frac{T}{T_r} \rceil$ intervals of duration T_r. Fix one such interval I. Let $S_I \triangleq \{(t,y) \in S : t \in I\}$, the set of points in S that correspond to times in I. Recall the notation $\Delta(x,y) \triangleq \mu^*(x) - \mu(x,y)$ and let

$$Y_I \triangleq \{y \in Y : \Delta(t_I, y) \leq 14\, r\}, \text{ where } t_I \triangleq \min(I). \tag{31}$$

All quantities in (31) refer to a fixed time t_I, which will allow us to use the uniform marginals property.

Note that Y_I contains at least one arm, namely the best arm $y^*(t_I)$. We claim that

$$|S_I| \leq 2\,|Y_I \setminus \{y^*(t_I)\}|. \tag{32}$$

Fix arm y. First, $\mathcal{D}_{\mathrm{X}}(t,t') \leq r$ for any $t,t' \in I$, so there exists at most one $t \in I$ such that $(t,y) \in S$. Second, suppose such t exists. Since $S \subset \mathcal{P}_{\mu,r}$, it follows that $\Delta(t,y) \leq 12\, r$. By Claim 18 it holds that

$$\Delta(t_I, y) \leq \Delta(t,y) + 2\,\mathcal{D}_{\mathrm{X}}(t,t') \leq 14\, r.$$

So $y \in Y_I$. It follows that $|S_I| \leq |Y_I|$.

To obtain (32), we show that $S_I = 0$ whenever $|Y_I| = 1$. Indeed, suppose $Y_I = \{y\}$ is a singleton set, and $|S_I| > 0$. Then $S_I = \{(t,y)\}$ for some $t \in I$. We will show that (t,y) is an r-winner, contradicting the definition of S. For any arm $y' \neq y$ and any time t' such that $\mathcal{D}_{\mathrm{X}}(t,t') \leq 2r$ it holds that

$$\begin{aligned}
\mu(t_I, y) &= \mu^*(t_I) > \mu(t_I, y') + 14r \\
\mu(t', y) &\geq \mu(t_I, y) - \mathcal{D}_{\mathrm{X}}(t', t_I) \geq \mu(t_I, y) - 3r \\
&> \mu(t_I, y') + 11r \\
&\geq \mu(t', y') - \mathcal{D}_{\mathrm{X}}(t', t_I) + 11r \\
&\geq \mu(t', y') + 8r.
\end{aligned}$$

and so $\mu(t',y) = \mu^*(t')$. Thus, (t,y) is an r-winner as claimed. This completes the proof of (32).

Now using (32) and Claim 19 we obtain that

$$\begin{aligned}
\mathbb{E}_\Gamma[\,|S_I|\,] &\leq 2\,\mathbb{E}_\Gamma[\,|Y_I \setminus \{y^*(t_I)\}|\,] \leq O(kr) \\
\mathbb{E}_\Gamma[\,|S|\,] &\leq \lceil \tfrac{T}{T_r} \rceil\, \mathbb{E}[\,|S_I|\,] \leq O(kr)\, \lceil \tfrac{T}{T_r} \rceil.
\end{aligned}$$

Taking the max over all possible S, we obtain $\mathbb{E}_\Gamma[\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r}] \leq O(kr)\ \lceil \frac{T}{T_r} \rceil$. To complete the proof of (30), we note that S cannot be r-consistent unless $|I| \geq 1/r^2$.

Now that we have (30), the rest is a simple computation. We use Theorem 6, namely we take (5) with $r_0 \to 0$, plug in (30), and recall that $T_r \geq 1/r^2 \iff r \geq \sigma^{1/3}$.

$$\begin{aligned} R(T) &\leq \textstyle\sum_{r=2^i \geq \sigma^{1/3}} \frac{1}{r}\ N_r\ O(\log T) \\ \mathbb{E}_\Gamma[R(T)] &\leq \textstyle\sum_{r=2^i \geq \sigma^{1/3}} O(k \log T)(\frac{T\sigma}{r} + 1) \\ &\leq O(k \log T)(T\sigma^{2/3} + \log \tfrac{1}{\sigma}). \end{aligned}$$

It follows that $\mathbb{E}_\Gamma[\hat{R}(T)] \leq O(k\,\sigma^{2/3} \log T)$ for any $T \geq \sigma^{-2/3} \log \frac{1}{\sigma}$. ■

Proof of Corollary 16 It suffices to bound $\hat{R}(T_0)$ for (non-periodic) contextual zooming.

Recall that expected payoffs satisfy the temporal constraint (26). Consider the high-probability event that

$$|\mu(t,y) - \mu(t',y)| \leq \sigma\,|t-t'|^{1/2} \log T_0 \qquad \forall t,t' \in [1,T_0],\ y \in Y. \tag{33}$$

Since expected regret due to the failure of (33) is negligible, from here on we will assume that (33) holds deterministically.

Let $\mathcal{D}_X(t,t') \triangleq \sigma\,|t-t'|^{1/2} \log T_0$ be the distance on contexts implicit in (33). For each $r > 0$, define $T_r \triangleq (\frac{r}{\sigma \log T_0})^2$. Then (30) follows exactly as in the proof of Corollary 15. We use Theorem 6 similarly: we take (5) with $r_0 \to 0$, plug in (30), and note that $T_r \geq 1/r^2 \iff r \geq (\sigma \log T_0)^{1/2}$. We obtain

$$\begin{aligned} \mathbb{E}_\Gamma[R(T_0)] &\leq \sum_{r=2^i \geq (\sigma \log T_0)^{1/2}} O(k \log T_0)(\tfrac{T_0}{T_r} + 1) \\ &\leq O(k \log^2 T_0)(T_0\,\sigma + \log \tfrac{1}{\sigma}). \end{aligned}$$

It follows that $\mathbb{E}_\Gamma[\hat{R}(T)] \leq O(k\,\sigma \log^2 T_0)$ as long as $T_0 \geq \frac{1}{\sigma} \log \frac{1}{\sigma}$. ■

8. Contextual Bandits with Adversarial Payoffs

In this section we consider the adversarial setting. We provide an algorithm which maintains an adaptive partition of the context space and thus takes advantage of "benign" context arrivals. It is in fact a *meta-algorithm*: given a bandit algorithm `Bandit`, we present a contextual bandit algorithm, called `ContextualBandit`, which calls `Bandit` as a subroutine.

8.1 Our Setting

Recall that in each round t, the context $x_t \in X$ is revealed, then the algorithm picks an arm $y_t \in Y$ and observes the payoff $\pi_t \in [0,1]$. Here X is the context set, and Y is the arms set. In this section, all context-arms pairs are feasible: $\mathcal{P} = X \times Y$.

Adversarial payoffs are defined as follows. For each round t, there is a payoff function $\hat{\pi}_t : X \times Y \to [0,1]$ such that $\pi_t = \hat{\pi}_t(x_t, y_t)$. The payoff function $\hat{\pi}_t$ is sampled independently

from a time-specific distribution Π_t over payoff functions. Distributions Π_t are fixed by the adversary in advance, before the first round, and not revealed to the algorithm. Denote $\mu_t(x,y) \triangleq \mathbb{E}[\Pi_t(x,y)]$.

Following Hazan and Megiddo (2007), we generalize the notion of regret for context-free adversarial MAB to contextual MAB. The context-specific best arm is

$$y^*(x) \in \text{argmax}_{y \in Y} \sum_{t=1}^{T} \mu_t(x,y), \tag{34}$$

where the ties are broken in an arbitrary but fixed way. We define *adversarial contextual regret* as

$$R(T) \triangleq \sum_{t=1}^{T} \mu_t(x_t,y_t) - \mu_t^*(x_t), \quad \text{where} \quad \mu_t^*(x) \triangleq \mu_t(x,y^*(x)). \tag{35}$$

Similarity information is given to an algorithm as a pair of metric spaces: a metric space $(X,\mathcal{D}_X)$ on contexts (the *context space*) and a metric space $(Y,\mathcal{D}_Y)$ on arms (the *arms space*), which form the product similarity space $(X\times Y, \mathcal{D}_X+\mathcal{D}_Y)$. We assume that for each round t functions μ_t and μ_t^* are Lipschitz on $(X\times Y,\mathcal{D}_X+\mathcal{D}_Y)$ and $(X,\mathcal{D}_X)$, respectively, both with Lipschitz constant 1 (see Footnote 1). We assume that the context space is compact, in order to ensure that the max in (34) is attained by some $y \in Y$. Without loss of generality, $\texttt{diameter}(X,\mathcal{D}_X) \leq 1$.

Formally, a problem instance consists of metric spaces $(X,\mathcal{D}_X)$ and $(Y,\mathcal{D}_Y)$, the sequence of context arrivals (denoted $x_{(1..T)}$), and a sequence of distributions $(\Pi_t)_{t\leq T}$. Note that for a fixed distribution $\Pi_t = \Pi$, this setting reduces to the stochastic setting, as defined in Introduction. For the fixed context case ($x_t = x$ for all t) this setting reduces to the (context-free) MAB problem with a randomized oblivious adversary.

8.2 Our Results

Our algorithm is parameterized by a regret guarantee for `Bandit` for the fixed context case, namely an upper bound on the convergence time.[14] For a more concrete theorem statement we will assume that the convergence time of `Bandit` is at most $T_0(r) \triangleq c_Y\, r^{-(2+d_Y)}\, \log(\frac{1}{r})$ for some constants c_Y and d_Y that are known to the algorithm. In particular, an algorithm in Kleinberg (2004) achieves this guarantee if d_Y is the c-covering dimension of the arms space and $c_Y = O(c^{2+d_Y})$.

This is a flexible formulation that can leverage prior work on adversarial bandits. For instance, if $Y \subset \mathbb{R}^d$ and for each fixed context $x \in X$ distributions Π_t randomize over linear functions $\hat{\pi}_t(x,\cdot) : Y \to \mathbb{R}$, then one could take `Bandit` from the line of work on adversarial bandits with linear payoffs. In particular, there exist algorithms with $d_Y = 0$ and $c_Y = \text{poly}(d)$ (Dani et al., 2007; Abernethy et al., 2008; Bubeck et al., 2012). Likewise, for convex payoffs there exist algorithms with $d_Y = 2$ and $c_Y = O(d)$ (Flaxman et al., 2005). For a bounded number of arms, algorithm EXP3 (Auer et al., 2002b) achieves $d_Y = 0$ and $c_Y = O(\sqrt{|Y|})$.

From here on, the context space $(X,\mathcal{D}_X)$ will be only metric space considered; balls and other notions will refer to the context space only.

14. The r-convergence time $T_0(r)$ is the smallest T_0 such that regret is $R(T) \leq rT$ for each $T \geq T_0$.

To quantify the "goodness" of context arrivals, our guarantees are in terms of the covering dimension of $x_{(1..T)}$ rather than that of the entire context space. (This is the improvement over the guarantee (3) for the uniform algorithm.) In fact, use a more refined notion which allows to disregard a limited number of "outliers" in $x_{(1..T)}$.

Definition 20 *Given a metric space and a multi-set S, the* (r,k)-covering number *of S is the r-covering number of the set $\{x \in S : |B(x,r) \cap S| \geq k\}$.*[15] *Given a constant c and a function $k : (0,1) \to \mathbb{N}$, the* ***relaxed covering dimension*** *of S with slack $k(\cdot)$ is the smallest $d > 0$ such that the $(r, k(r))$-covering number of S is at most $c\, r^{-d}$ for all $r > 0$.*

Our result is stated as follows:

Theorem 21 *Consider the contextual MAB problem with adversarial payoffs, and let* `Bandit` *be a bandit algorithm. Assume that the problem instance belongs to some class of problem instances such that for the fixed-context case, convergence time of* `Bandit` *is at most $T_0(r) \triangleq c_Y\, r^{-(2+d_Y)} \log(\frac{1}{r})$ for some constants c_Y and d_Y that are known to the algorithm. Then* `ContextualBandit` *achieves adversarial contextual regret $R(\cdot)$ such that for any time T and any constant $c_X > 0$ it holds that*

$$R(T) \leq O(c_{\text{DBL}}^2\, (c_X\, c_Y)^{1/(2+d_X+d_Y)})\ T^{1-1/(2+d_X+d_Y)}(\log T), \tag{36}$$

where d_X is the relaxed covering dimension of $x_{(1..T)}$ with multiplier c_X and slack $T_0(\cdot)$, and c_{DBL} is the doubling constant of $x_{(1..T)}$.

Remarks. For a version of (36) that is stated in terms of the "raw" (r, k_r)-covering numbers of $x_{(1..T)}$, see (38) in the analysis (page 2563).

8.3 Our Algorithm

The contextual bandit algorithm `ContextualBandit` is parameterized by a (context-free) bandit algorithm `Bandit`, which it uses as a subroutine, and a function $T_0(\cdot) : (0,1) \to \mathbb{N}$.

The algorithm maintains a finite collection $\mathcal{A}$ of balls, called *active balls*. Initially there is one active ball of radius 1. Ball B stays active once it is *activated*. Then a fresh instance $\texttt{ALG}_B$ of `Bandit` is created, whose set of "arms" is Y. $\texttt{ALG}_B$ can be parameterized by the time horizon $T_0(r)$, where r is the radius of B.

The algorithm proceeds as follows. In each round t the algorithm selects one active ball $B \in \mathcal{A}$ such that $x_t \in B$, calls $\texttt{ALG}_B$ to select an arm $y \in Y$ to be played, and reports the payoff π_t back to $\texttt{ALG}_B$. A given ball can be selected at most $T_0(r)$ times, after which it is called *full*. B is called *relevant* in round t if it contains x_t and is not full. The algorithm selects a relevant ball (breaking ties arbitrarily) if such ball exists. Otherwise, a new ball B' is activated and selected. Specifically, let B be the smallest-radius active ball containing x_t. Then $B' = B(x_t, \frac{r}{2})$, where r is the radius of B. B is then called the *parent* of B'. See Algorithm 2 for the pseudocode.

15. By abuse of notation, here $|B(x,r) \cap S|$ denotes the number of points $x \in S$, with multiplicities, that lie in $B(x,r)$.

Algorithm 2 Algorithm `ContextualBandit`.

Input:
  Context space $(X, \mathcal{D}_X)$ of diameter ≤ 1, set Y of arms.
  Bandit algorithm `Bandit` and a function $T_0(\cdot) : (0,1) \to \mathbb{N}$.
Data structures:
  A collection $\mathcal{A}$ of "active balls" in $(X, \mathcal{D}_X)$.
  $\forall B \in \mathcal{A}$: counter n_B, instance $\texttt{ALG}_B$ of `Bandit` on arms Y.
Initialization: $B \leftarrow B(x, 1)$; // center $x \in X$ is arbitrary
  $\mathcal{A} \leftarrow \{B\}$; $n_B \leftarrow 0$; initiate $\texttt{ALG}_B$.
  $\mathcal{A}^* \leftarrow \mathcal{A}$ // active balls that are not full
Main loop: for each round t
  Input context x_t.
  $\texttt{relevant} \leftarrow \{B \in \mathcal{A}^* : x_t \in B\}$.
  if $\texttt{relevant} \neq \emptyset$ **then**
    $B \leftarrow$ any $B \in \texttt{relevant}$.
  else // activate a new ball:
    $r \leftarrow \min_{B \in \mathcal{A}:\, x_t \in B} r_B$.
    $B \leftarrow B(x_t, r/2)$. // new ball to be added
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{B\}$; $\mathcal{A}^* \leftarrow \mathcal{A}^* \cup \{B\}$; $n_B \leftarrow 0$; initiate $\texttt{ALG}_B$.
  $y \leftarrow$ next arm selected by $\texttt{ALG}_B$.
  Play arm y, observe payoff π, report π to $\texttt{ALG}_B$.
  $n_B \leftarrow n_B + 1$.
  if $n_B = T_0(\texttt{radius}(B))$ **then** $\mathcal{A}^* \leftarrow \mathcal{A}^* \setminus \{B\}$. // ball B is full

8.4 Analysis: Proof of Theorem 21

First let us argue that algorithm `ContextualBandit` is well-defined. Specifically, we need to show that after the activation rule is called, there exists an active non-full ball containing x_t. Suppose not. Then the ball $B' = B(x_t, \frac{r}{2})$ activated by the activation rule must be full. In particular, B' must have been active before the activation rule was called, which contradicts the minimality in the choice of r. Claim proved.

We continue by listing several basic claims about the algorithm.

Claim 22 *The algorithm satisfies the following basic properties:*

(a) (Correctness) In each round t, exactly one active ball is selected.

(b) Each active ball of radius r is selected at most $T_0(r)$ times.

(c) (Separation) For any two active balls $B(x, r)$ and $B(x', r)$ we have $\mathcal{D}_X(x, x') > r$.

(d) Each active ball has at most c_{DBL}^2 children, where c_{DBL} is the doubling constant of $x_{(1..T)}$.

Proof Part (a) is immediate from the algorithm's specification. For (b), simply note that by the algorithms' specification a ball is selected only when it is not full.

To prove (c), suppose that $\mathcal{D}_X(x, x') \leq r$ and suppose $B(x', r)$ is activated in some round t while $B(x, r)$ is active. Then $B(x', r)$ was activated as a child of some ball B^* of radius $2r$. On the other hand, $x' = x_t \in B(x, r)$, so $B(x, r)$ must have been full in round t (else no ball would have been activated), and consequently the radius of B^* is at most r. Contradiction.

For (d), consider the children of a given active ball $B(x,r)$. Note that by the activation rule the centers of these children are points in $x_{(1..T)} \cap B(x,r)$, and by the separation property any two of these points lie within distance $> \frac{r}{2}$ from one another. By the doubling property, there can be at most c_{DBL}^2 such points. ■

Let us fix the time horizon T, and let $R(T)$ denote the contextual regret of `ContextualBandit`. Partition $R(T)$ into the contributions of active balls as follows. Let $\mathcal{B}$ be the set of all balls that are active after round T. For each $B \in \mathcal{B}$, let S_B be the set of all rounds t when B has been selected. Then

$$R(T) = \textstyle\sum_{B\in\mathcal{B}} R_B(T), \quad \text{where} \quad R_B(T) \triangleq \sum_{t\in S_B} \mu_t^*(x_t) - \mu_t(x_t,\, y_t).$$

Claim 23 *For each ball $B = B(x,r) \in \mathcal{B}$, we have $R_B \leq 3\, r\, T_0(r)$.*

Proof By the Lipschitz conditions on μ_t and μ_t^*, for each round $t \in S_B$ it is the case that

$$\mu_t^*(x_t) \leq r + \mu_t^*(x) = r + \mu_t(x, y^*(x)) \leq 2rn + \mu_t(x_t,\, y^*(x)).$$

The t-round regret of `Bandit` is at most $R_0(t) \triangleq t\, T_0^{-1}(t)$. Therefore, letting $n = |S_B|$ be the number of times algorithm $\texttt{ALG}_B$ has been invoked, we have that

$$R_0(n) + \textstyle\sum_{t\in S_B} \mu_t(x_t,\, y_t) \geq \sum_{t\in S_B} \mu_t(x_t, y^*(x)) \geq \sum_{t\in S_B} \mu_t^*(x_t) - 2rn.$$

Therefore $R_B(T) \leq R_0(n) + 2rn$. Recall that by Claim 22(b) we have $n \leq T_0(r)$. Thus, by definition of convergence time $R_0(n) \leq R_0(T_0(r)) \leq r\, T_0(r)$, and therefore $R_B(T) \leq 3\, r\, T_0(r)$. ■

Let $\mathcal{F}_r$ be the collection of all full balls of radius r. Let us bound $|\mathcal{F}_r|$ in terms the (r,k)-covering number of $x_{(1..T)}$ in the context space, which we denote $N(r,k)$.

Claim 24 *There are at most $N(r,\, T_0(r))$ full balls of radius r.*

Proof Fix r and let $k = T_0(r)$. Let us say that a point $x \in x_{(1..T)}$ is *heavy* if $B(x,r)$ contains at least k points of $x_{(1..T)}$, counting multiplicities. Clearly, $B(x,r)$ is full only if its center is heavy. By definition of the (r,k)-covering number, there exists a family $\mathcal{S}$ of $N(r,k)$ sets of diameter $\leq r$ that cover all heavy points in $x_{(1..T)}$. For each full ball $B = B(x,r)$, let S_B be some set in $\mathcal{S}$ that contains x. By Claim 22(c), the sets S_B, $B \in \mathcal{F}_r$ are all distinct. Thus, $|\mathcal{F}_r| \leq |\mathcal{S}| \leq N(r,k)$. ■

Let $\mathcal{B}_r$ be the set of all balls of radius r that are active after round T. By the algorithm's specification, each ball in $\mathcal{F}_r$ has been selected $T_0(r)$ times, so $|\mathcal{F}_r| \leq T/T_0(r)$. Then using Claim 22(b) and Claim 24, we have

$$\begin{aligned} |\mathcal{B}_{r/2}| &\leq c_{\text{DBL}}^2\, |\mathcal{F}_r| \leq c_{\text{DBL}}^2\ \min(T/T_0(r),\ N(r,\, T_0(r))) \\ \textstyle\sum_{B\in\mathcal{B}_{r/2}} R_B &\leq O(r)\, T_0(r)\, |\mathcal{B}_{r/2}| \leq O(c_{\text{DBL}}^2)\ \min(rT,\ r\, T_0(r)\, N(r, T_0(r))). \end{aligned} \tag{37}$$

Trivially, for any full ball of radius r we have $T_0(r) \leq T$. Thus, summing (37) over all such r, we obtain

$$R(T) \leq O(c_{\text{DBL}}^2) \sum_{r=2^{-i}:\, i\in\mathbb{N} \text{ and } T_0(r)\leq T} \min(rT,\ r\,T_0(r)\,N(r,T_0(r))). \tag{38}$$

Note that (38) makes no assumptions on $N(r,T_0(r))$. Now, plugging in $T_0(r) = c_{\text{Y}}\, r^{-(2+d_{\text{Y}})}$ and $N(r,T_0(r)) \leq c_{\text{X}}\, r^{-d_{\text{X}}}$ into (38) and optimizing it for r it is easy to derive the desired bound (36).

9. Conclusions

We consider a general setting for contextual bandit problems where the algorithm is given information on similarity between the context-arm pairs. The similarity information is modeled as a metric space with respect to which expected payoffs are Lipschitz-continuous. Our key contribution is an algorithm which maintains a partition of the metric space and adaptively refines this partition over time. Due to this "adaptive partition" technique, one can take advantage of "benign" problem instances without sacrificing the worst-case performance; here "benign-ness" refers to both expected payoffs and context arrivals. We essentially resolve the setting where expected payoff from every given context-arm pair either does not change over time, or changes slowly. In particular, we obtain nearly matching lower bounds (for time-invariant expected payoffs and for an important special case of slow change).

We also consider the setting of adversarial payoffs. For this setting, we design a different algorithm that maintains a partition of contexts and adaptively refines it so as to take advantage of "benign" context arrivals (but not "benign" expected payoffs), without sacrificing the worst-case performance. Our algorithm can work with, essentially, any given off-the-shelf algorithm for standard (non-contextual) bandits, the choice of which can then be tailored to the setting at hand.

The main open questions concern relaxing the requirements on the quality of similarity information that are needed for the provable guarantees. First, it would be desirable to obtain similar results under weaker versions of the Lipschitz condition. Prior work (Kleinberg et al., 2008b; Bubeck et al., 2011a) obtained several such results for the non-contextual version of the problem, mainly because their main results do not require the full power of the Lipschitz condition. However, the analysis in this paper appears to make a heavier use of the Lipschitz condition; it is not clear whether a meaningful relaxation would suffice. Second, in some settings the available similarity information might not include any numeric upper bounds on the difference in expected payoffs; e.g., it could be given as a tree-based taxonomy on context-arm pairs, without any explicit numbers. Yet, one wants to recover the same provable guarantees *as if* the numerical information were explicitly given. For the non-contextual version, this direction has been explored in (Bubeck et al., 2011b; Slivkins, 2011).[16]

Another open question concerns our results for adversarial payoffs. Here it is desirable to extend our "adaptive partitions" technique to also take advantage of "benign" expected

16. (Bubeck et al., 2011b; Slivkins, 2011) have been published after the preliminary publication of this paper on `arxiv.org`.

payoffs (in addition to "benign" context arrivals). However, to the best of our knowledge such results are not even known for the non-contextual version of the problem.

Acknowledgments

The author is grateful to Ittai Abraham, Bobby Kleinberg and Eli Upfal for many conversations about multi-armed bandits, and to Sebastien Bubeck for help with the manuscript. Also, comments from anonymous COLT reviewers and JMLR referees have been tremendously useful in improving the presentation.

Appendix A. The KL-divergence Technique, Encapsulated

To analyze the lower-bounding construction in Section 5, we use an extension of the KL-divergence technique from Auer et al. (2002b), which is implicit in Kleinberg (2004) and encapsulated as a stand-alone theorem in Kleinberg et al. (2013). To make the paper self-contained, we state the theorem from Kleinberg et al. (2013), along with the relevant definitions. The remainder of this section is copied from Kleinberg et al. (2013), with minor modifications.

Consider a very general MAB setting where the algorithm is given a strategy set X and a collection $\mathcal{F}$ of feasible payoff functions; we call it the *feasible MAB problem* on $(X, \mathcal{F})$. For example, $\mathcal{F}$ can consist of all functions $\mu : X \to [0, 1]$ that are Lipschitz with respect to a given metric space. The lower bound relies on the existence of a collection of subsets of $\mathcal{F}$ with certain properties, as defined below. These subsets correspond to children of a given tree node in the ball-tree

Definition 25 *Let X be the strategy set and $\mathcal{F}$ be the set of all feasible payoff functions. An* (ϵ, k)-ensemble *is a collection of subsets $\mathcal{F}_1, \ldots, \mathcal{F}_k \subset \mathcal{F}$ such that there exist mutually disjoint subsets $S_1, \ldots, S_k \subset X$ and a number $\mu_0 \in [\frac{1}{3}, \frac{2}{3}]$ which satisfy the following. Let $S = \cup_{i=1}^{k} S_i$. Then*

- *on $X \setminus S$, any two functions in $\cup_i \mathcal{F}_i$ coincide, and are bounded from above by μ_0.*
- *for each i and each function $\mu \in \mathcal{F}_i$ it holds that $\mu = \mu_0$ on $S \setminus S_i$ and $\sup(\mu_i, S_i) = \mu_0 + \epsilon$.*

Assume the payoff function μ lies in $\cup_i \mathcal{F}_i$. The idea is that an algorithm needs to play arms in S_i for at least $\Omega(\epsilon^{-2})$ rounds in order to determine whether $\mu \in \mathcal{F}_i$, and each such step incurs ϵ regret if $\mu \notin \mathcal{F}_i$. In our application, subsets $S_1, \ldots, S_k$ correspond to children $u_1, \ldots, u_k$ of a given tree node in the ball-tree, and each $\mathcal{F}_i$ consists of payoff functions induced by the ends in the subtree rooted at u_i.

Theorem 26 (Theorem 5.6 in Kleinberg et al. (2013)) *Consider the feasible MAB problem with 0-1 payoffs. Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be an (ϵ, k)-ensemble, where $k \geq 2$ and $\epsilon \in (0, \frac{1}{12})$. Then for any $t \leq \frac{1}{32} k \epsilon^{-2}$ and any bandit algorithm there exist at least $k/2$ distinct i's such that the regret of this algorithm on any payoff function from $\mathcal{F}_i$ is at least $\frac{1}{60} \epsilon t$.*

In Auer et al. (2002b), the authors analyzed a special case of an (ϵ, k)-ensemble in which there are k arms $u_1, \ldots, u_k$, and each $\mathcal{F}_i$ consists of a single payoff function that assigns expected payoff $\frac{1}{2} + \epsilon$ to arm u_i, and $\frac{1}{2}$ to all other arms.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21th Conf. on Learning Theory (COLT)*, pages 263–274, 2008.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, 1995.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. of Machine Learning Research (JMLR)*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. Preliminary version in *15th ICML*, 1998.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *20th Conf. on Learning Theory (COLT)*, pages 454–468, 2007.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version in *36th ACM STOC*, 2004.

Jeffrey Banks and Rangarajan Sundaram. Denumerable-armed bandits. *Econometrica*, 60 (5):1071–1096, 1992.

Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *Annals of Statistics*, 25(5):2103–2116, 1997.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck and Rémi Munos. Open loop optimistic planning. In *23rd Conf. on Learning Theory (COLT)*, pages 477–489, 2010.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. Online optimization in X-armed bandits. *J. of Machine Learning Research (JMLR)*, 12:1587–1627, 2011a. Preliminary version in *NIPS 2008*.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the Lipschitz constant. In *22nd Intl. Conf. on Algorithmic Learning Theory (ALT)*, pages 144–158, 2011b.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *25th Conf. on Learning Theory (COLT)*, 2012.

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

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Richard Cole and Lee-Ad Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *38th ACM Symp. on Theory of Computing (STOC)*, pages 574–583, 2006.

Varsha Dani, Thomas P. Hayes, and Sham Kakade. The price of bandit information for online optimization. In *20th Advances in Neural Information Processing Systems (NIPS)*, 2007.

Abraham Flaxman, Adam Kalai, and Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *16th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 385–394, 2005.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low–distortion embeddings. In *44th IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 534–543, 2003.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 38–47, 2009.

Elad Hazan and Nimrod Megiddo. Online learning with prior information. In *20th Conf. on Learning Theory (COLT)*, pages 499–513, 2007.

Juha Heinonen. *Lectures on Analysis on Metric Spaces*. Universitext. Springer-Verlag, New York, 2001.

Jon Kleinberg, Aleksandrs Slivkins, and Tom Wexler. Triangulation and embedding using small sets of beacons. In *45th IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 444–453, 2004.

Jon Kleinberg, Aleksandrs Slivkins, and Tom Wexler. Triangulation and embedding using small sets of beacons. *J. of the ACM*, 56(6), September 2009. Subsumes conference papers in *IEEE FOCS 2004* Kleinberg et al. (2004) and *ACM-SIAM SODA 2005* Slivkins (2005).

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems (NIPS)*, 2004.

Robert Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, MIT, 2005.

Robert Kleinberg and Aleksandrs Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *21st ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2010.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *21st Conf. on Learning Theory (COLT)*, pages 425–436, 2008a.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pages 681–690, 2008b.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. Technical report `http://arxiv.org/abs/1312.1277`. Merged and revised version of conference papers in *ACM STOC 2008* and *ACM-SIAM SODA 2010*, Dec 2013.

Levente Kocsis and Csaba Szepesvari. Bandit based Monte-Carlo planning. In *17th European Conf. on Machine Learning (ECML)*, pages 282–293, 2006.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *21st Advances in Neural Information Processing Systems (NIPS)*, 2007.

Alessandro Lazaric and Rémi Munos. Hybrid stochastic-adversarial on-line learning. In *22nd Conf. on Learning Theory (COLT)*, 2009.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th Intl. World Wide Web Conf. (WWW)*, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *4th ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, 2011.

Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via Lipschitz context multi-armed bandits. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Odalric-Ambrym Maillard and Rémi Munos. Online learning in adversarial Lipschitz environments. In *European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 305–320, 2010.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *22nd Conf. on Learning Theory (COLT)*, 2009.

Rémi Munos and Pierre-Arnaud Coquelin. Bandit algorithms for tree search. In *23rd Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2007.

Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *25th Intl. Conf. on Machine Learning (ICML)*, pages 784–791, 2008.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. In *23rd Conf. on Learning Theory (COLT)*, pages 54–66, 2010.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

Aleksandrs Slivkins. Distributed approaches to triangulation and embedding. In *16th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 640–649, 2005.

Aleksandrs Slivkins. Multi-armed bandits on implicit metric spaces. In *25th Advances in Neural Information Processing Systems (NIPS)*, 2011.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the Brownian restless bandits. In *21st Conf. on Learning Theory (COLT)*, pages 343–354, 2008.

Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Learning optimally diverse rankings over large document collections. *J. of Machine Learning Research (JMLR)*, 14 (Feb):399–436, 2013. Preliminary version in *27th ICML*, 2010.

Kunal Talwar. Bypassing the embedding: Algorithms for low-dimensional metrics. In *36th ACM Symp. on Theory of Computing (STOC)*, pages 281–290, 2004.

Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Trans. on Automatic Control*, 50(3):338355, 2005.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1729–1736, 2008.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.*, 74(368), 1979.