# Contextual Bandits with Similarity Information[*]

**Aleksandrs Slivkins**
Microsoft Research Silicon Valley
slivkins@microsoft.com

## 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 webpages, 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).

## 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 1952 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.

[*] The full version of this paper is available from arxiv.org. A preliminary version (which does not include the results in Section 6) has been posted to arxiv.org in July 2009. The results in Section 7 have been obtained while the author was a postdoc at Brown University, and a write-up has been circulated in 2007.

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 webpages, 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.

**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 $\pi_t$ 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 7).

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*:

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

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* $(\mathcal{P}, \mathcal{D}) = (X \times Y, \mathcal{D}_\mathrm{X} + \mathcal{D}_\mathrm{Y})$, 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

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

**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 $S_\mathrm{X}$ of the context space and approximates $x_t$ by the closest point in $S_\mathrm{X}$, 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 $S_\mathrm{Y}$ of the arms space, and run a $k$-arm bandit algorithm such as EXP3 Auer et al. (2002b) on the points in $S_\mathrm{Y}$. 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 $d_\mathrm{X}$ is the covering dimension of the context space and $d_\mathrm{Y}$ is that of the arms space.

[1] In other words, $\mu$ 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 $K_\mathrm{Lip}$ is known to the algorithm), since we can re-define $\mathcal{D} \leftarrow K_\mathrm{Lip}\, D$.

**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 techniques from (Auer et al., 2002b, Kleinberg, 2004, Kleinberg et al., 2008b). 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. (2010).

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.

**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., 2008). 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 (see Discussion 4.1). 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.

**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 7. All omitted proofs appear in the full version.

## 2 Related work

A proper discussion of the literature on bandit problems is beyond the scope of this paper. A reader is encouraged to refer to Cesa-Bianchi and Lugosi (2006) for background.

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., 2008, 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). 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. (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. In particular, Maillard and Munos (2010) achieve regret $\tilde{O}(\sqrt{dT})$ for $d$-dimensional real space, whereas (even) for stochastic payoffs there is a lower bound $\Omega(T^{1-1/(d+2)})$ (Kleinberg, 2004, Bubeck et al., 2008).

**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. 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 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.

**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* of $\mathcal{P}$ and denoted $N_r(\mathcal{P})$. [3] 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* $c_{\mathrm{DBL}}(\mathcal{P})$ of $\mathcal{P}$ is the smallest $k$ such that any ball can be covered by $k$ balls of half the radius. The doubling constant has been a standard notion in theoretical computer science since Gupta et al. (2003). It is known that that $c_{\mathrm{DBL}}(\mathcal{P})\geq c\,2^d$ if $d$ is the covering dimension of $\mathcal{P}$ with multiplier $c$, and that $c_{\mathrm{DBL}}(\mathcal{P})\leq 2^d$ if $\mathcal{P}$ is a 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 $c_{\mathrm{DBL}}^2$ points of $S$.

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

[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.

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 $K_{\text{Lip}}$, 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 $\gamma>0$.

# 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.

**Discussion 4.1.** 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 4.4 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, esp. 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 $\{N_r\}_{r\in(0,1)}$:

$$R(T)\leq C_0\,\inf_{r_0\in(0,1)}\left(r_0T+\textstyle\sum_{r=2^{-i}:\,i\in\mathbb{N},\,r_0\leq r\leq 1}\;\frac{1}{r}\,N_r\log T\right). \tag{5}$$

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 of $\mathcal{P}$. [4] 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. (2008)) 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 $d_{\text{X}}+d_{\text{Y}}$), 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:

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

[4] Then (5) can be simplified to $R(T)\leq\inf_{r\in(0,1)}O\left(rT+\frac{1}{r}\,N_r(\mathcal{P})\log T\right)$ since $N_r(\mathcal{P})$ 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 $c$).

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 4.2.** *Consider the contextual MAB problem with stochastic payoffs. The contextual regret $R(T)$ of the contextual zooming algorithm satisfies (5), where $N_r = N_r(\mathcal{P}_{\mu,r})$ is the $r$-zooming number. Consequently, $R(T)$ satisfies the dimension-type guarantee (7), where $d_c$ is the contextual zooming dimension.*

In Theorem 4.2, 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.

On 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".

Let us define the two rules. First we need 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

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

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 $t$, denoted $\texttt{dom}\,(B, \mathcal{A}_t)$, is a subset of $B$ that excludes all balls $B' \in \mathcal{A}_t$ of strictly smaller radius:

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

$B$ is called *relevant* in round $t$ if $(x_t, y) \in \texttt{dom}\,(B, \mathcal{A}_t)$ for some arm $y$. In each round, the algorithm chooses among relevant balls $B$ 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:

- ***selection rule.*** In round $t$, select a relevant ball $B$ with the maximal index (break ties arbitrarily). Select an arbitrary arm $y$ such that $(x_t, y) \in \texttt{dom}\,(B, \mathcal{A}_t)$.
- ***activation rule.*** If in round $t$ the selection rule selects $(B, y)$ such that $\texttt{rad}(n_t(B) + 1) \leq r(B)$, then 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.)

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":

$$I_t^{\text{pre}}(B) \triangleq \nu_t(B) + 2\, r(B) + \texttt{rad}_t(B). \tag{11}$$

The "uncertainty term" in (11) reflects both uncertainty due to a location in the metric space and uncertainty due to an insufficient number of samples. The index of $B$ is obtained by taking a minimum over all active balls $B'$ of radius at least $r(B)$ (letting $\mathcal{D}(B, B')$ is the distance between the centers of the two balls).

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

### 4.3 Analysis of the algorithm: proof of Theorem 4.2

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

**Claim 4.3.** *The following invariants are maintained:*

- *(centering) if $B$ is activated in round $t$ with parent $B^{par}$, then the center of $B$ is $(x_t, y_t) \in \texttt{dom}\,(B^{par}, \mathcal{A})$.*
- *(confidence) $\texttt{rad}_t(B) > r(B)$ for all active balls $B$ and all rounds $t$.*
- *(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 at least $r$.*

*Proof.* The first two invariants are immediate. For the covering invariant, note that $\cup_{B\in\mathcal{A}} \mathtt{dom}\,(B,\mathcal{A}) = \cup_{B\in\mathcal{A}} B$ 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 $v \in \mathtt{dom}\,(B,\mathcal{A})$.) 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 $B^{\text{par}}$, and $B'$ is activated before time $t$. The center of $B$ is some point $(x_t, y_t) \in \mathtt{dom}\,(B^{\text{par}}, \mathcal{A}_t)$. 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 $B_t^{\mathtt{sel}}$ 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 4.4.** *If ball $B$ is active in round $t$, then with probability at least $1 - T^{-2}$ we have that*

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

*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 $\frac{1}{k}\,|Z_{k\wedge n}| \leq \mathtt{rad}_t(B)$, for each $k \leq T$. Taking the Union Bound, it follows that $\frac{1}{n}\,|Z_n| \leq \mathtt{rad}_t(B)$. 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. □

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 4.5.** *Consider a clean run of the algorithm. Then $\Delta(x_t, y_t) \leq 15\, r(B_t^{\mathtt{sel}})$ 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 $I_t(B) = I^{\text{pre}}(B') + \mathcal{D}(B,B')$ for some active ball $B'$ of radius $r(B') \geq r(B)$. Therefore

$$\begin{aligned}
I_t(B_t^{\mathtt{sel}}) &\geq I_t(B) = I^{\text{pre}}(B') + \mathcal{D}(B,B') && \text{(selection rule, defn of index (12))}\\
&= \nu_t(B') + 2\,r(B') + \mathtt{rad}_t(B') + \mathcal{D}(B,B') && \text{(defn of preindex (11))}\\
&\geq \mu(B') + r(B) + \mathcal{D}(B,B') && \text{(Claim 4.4 and } r(B') \geq r(B))\\
&\geq \mu(B) + r(B) \geq \mu(x_t,\, y^*(x_t)) = \mu^*(x_t). && \text{(Lipschitz property (1), twice)}
\end{aligned} \tag{14}$$

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

$$\mathtt{rad}_t(B^{\text{par}}) \leq r(B^{\text{par}}) = 2\,r(B_t^{\mathtt{sel}}), \tag{15}$$

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

$$\begin{aligned}
I_t(B_t^{\mathtt{sel}}) &\leq I^{\text{pre}}(B^{\text{par}}) + r(B^{\text{par}}) && \text{(defn of index (12))}\\
&= \nu_t(B^{\text{par}}) + 3\,r(B^{\text{par}}) + \mathtt{rad}_t(B^{\text{par}}) && \text{(defn of preindex (11))}\\
&\leq \mu(B^{\text{par}}) + 4\,r(B^{\text{par}}) + 2\,\mathtt{rad}_t(B^{\text{par}}) && \text{(Claim 4.4)}\\
&\leq \mu(B^{\text{par}}) + 12\,r(B_t^{\mathtt{sel}}) && \text{("parenthood" (15))}\\
&\leq \mu(x_t,y_t) + 15\,r(B_t^{\mathtt{sel}}) && \text{(Lipschitz property (1)).}
\end{aligned} \tag{16}$$

In the last inequality we used the fact that $(x_t,y_t)$ is within distance $3\,r(B_t^{\mathtt{sel}})$ from the center of $B^{\text{par}}$. Putting the pieces together, $\mu^*(x_t) \leq I_t(B_t^{\mathtt{sel}}) \leq \mu(x_t,y_t) + 15\,r(B_t^{\mathtt{sel}})$. □

**Corollary 4.6.** *In a clean run, if ball $B$ is activated in round $t$ then $\Delta(x_t,y_t) \leq 12\,r(B)$.*

*Proof.* By the activation rule, $B_t^{\mathtt{sel}}$ is the parent of $B$. Thus by Lemma 4.5 we immediately have $\Delta(x_t,y_t) \leq 15\,r(B_t^{\mathtt{sel}}) = 30\,r(B)$. To obtain the constant of 12 that is claimed here, it suffices to prove a more efficient special case of Lemma 4.5: if $\mathtt{rad}_t(B_t^{\mathtt{sel}}) \leq r(B_t^{\mathtt{sel}})$ then $\Delta(x_t,y_t) \leq 6\,r(B_t^{\mathtt{sel}})$. To prove this, we simply replace (16) in the proof of Lemma 4.5 by similar inequality in terms of $I^{\text{pre}}(B_t^{\mathtt{sel}})$ rather than $I^{\text{pre}}(B^{\text{par}})$:

$$\begin{aligned}
I_t(B_t^{\mathtt{sel}}) &\leq I^{\text{pre}}(B_t^{\mathtt{sel}}) = \nu_t(B_t^{\mathtt{sel}}) + 2\,r(B_t^{\mathtt{sel}}) + \mathtt{rad}_t(B_t^{\mathtt{sel}}) && \text{(defns (11-12))}\\
&\leq \mu(B_t^{\mathtt{sel}}) + 3\,r(B_t^{\mathtt{sel}}) + 2\,\mathtt{rad}_t(B_t^{\mathtt{sel}}) && \text{(Claim 4.4)}\\
&\leq \mu(x_t,y_t) + 6\,r(B_t^{\mathtt{sel}}) && \square
\end{aligned}$$

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. A ball $B \in \mathcal{F}_r$ is called *full* in round $t$ if $\mathtt{rad}_t(B) \leq r$. Note that in each round, if a full ball is selected then some other ball is activated. Thus, we will 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 not full. It is easy to see that $|S_B| \leq O(r^{-2}\,\log T)$. Moreover, by Lemma 4.5 and Corollary 4.6 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 by the activation rule its center is $(x_t, y_t)$, and Corollary 4.6 asserts that $(x_t, y_t) \in \mathcal{P}_{\mu,r}$, 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 4.2 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 4.2 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 $\mathtt{rad}_t(B) \leq r$ 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 context arrival $x_t$ can be assigned to exactly one point $p \in S$ so that $(x_t, y) \in B(p, r)$ for some arm $y$, and each point in $S$ is assigned at least $1/r^2$ arrivals. 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. 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 selected in a given round, then this round does not contribute to contextual regret. Therefore, it suffices to consider ($r$-consistent) $r$-packings of $\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r}$. This can be a significant saving if for most context arrivals $x_t$ expected payoff $\mu(x_t, y)$ is either optimal or very suboptimal (see Section 6.2 for an example).

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 4.7.** *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 $C_0 = O(c_{\text{DBL}})$. Here $c_{\text{DBL}}$ is the doubling constant of the similarity space. 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 4.2 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 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 $0$.

To formulate our result, let $R_\mu^{\text{UB}}(T)$ denote the upper bound in Theorem 4.2, 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 5.1.** *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) $R_\mu^{UB}(T) \leq O(R)$ 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. Roughly, for suitably chosen parameter $r \in (0,1)$ such that $N = T\,r^2 \leq N_r(\mathcal{P})$ we pick an $r$-net $S_{\text{X}}$ in the context space and an $r$-net $S_{\text{Y}}$ in the arms space so that $|S_{\text{X}}| \times |S_{\text{Y}}| = N$. For each $x \in S_{\text{X}}$ we construct a needle-in-the-haystack example on the set $S_{\text{Y}}$: we pick some $y^*(x) \in S_{\text{Y}}$ to be the "needle" (independently and uniformly at random), and define $\mu(x, y^*(x)) = \frac{1}{2} + \frac{r}{2}$, and $\mu(x,y) = \frac{1}{2} + \frac{r}{4}$ for other $y \in S_{\text{Y}}$. We smoothen the expected payoffs so that far from $S_{\text{X}} \times S_{\text{Y}}$ expected payoffs are $\frac{1}{2}$ and the Lipschitz condition holds. The sequence $x_{(1..T)}$ of context arrivals is defined in an arbitrary round-robin fashion over the points in $S_{\text{X}}$. We show that in $T$ rounds each context in $S_{\text{X}}$ contributes $\Omega(|S_{\text{Y}}|/r)$ to contextual regret resulting in total contextual regret $\Omega(N/r)$, and $R_\mu^{\text{UB}}(T) \leq O(N/r)(\log T)$ for each problem instance in our construction. See the full version for details.

# 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 6.3 we discuss a recent application of contextual zooming to bandit learning-to-rank, which has been published in Slivkins et al. (2010). Most of the proofs are deferred to the full version.

## 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 $\sigma_y$ in each round, for some a-priori known *volatilities* $\sigma_y$. The algorithm's goal here is to adapt to the changing environment. Thus, we define *dynamic regret*: regret with respect to a benchmark which in each round plays the best arm for this round. We are primarily interested in the long-term performance quantified by average dynamic regret $\hat{R}(T) \triangleq R(T)/T$. We call this setting the ***drifting MAB problem***.

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":

$$|\mu(t,y) - \mu(t',y)| \leq \sigma_y\,|t - t'| \tag{17}$$

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

$$\mathcal{D}((t,y),\,(t',y')) = \min(1,\; \sigma_y\,|t-t'| + \mathbf{1}_{\{y \neq y'\}}). \tag{18}$$

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* $T_0$. The general guarantees are provided by Theorem 4.2 and Theorem 4.7. Below we work out some specific, tractable corollaries.

**Corollary 6.1.** *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 $T \geq T_0 \geq (\frac{k}{\sigma^2})^{1/3} \log \frac{k}{\sigma}$.*

*Proof.* Since $\hat{R}(T) \leq 2\,\hat{R}(T_0)$ for any $T \geq T_0$, it suffices to bound $\hat{R}(T_0)$. Therefore, from here on we can focus on analyzing contextual zooming itself (rather than contextual zooming with period).

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

$$\text{dynamic regret } R(\cdot) \text{ satisfies (5) with } N_r \leq k\,\lceil \tfrac{T\sigma}{r} \rceil. \tag{19}$$

Plugging $N_r \leq k\,(1+\frac{T\sigma}{r})$ into (5) and taking $r_0 = (k\sigma \log T)^{1/3}$ we obtain[6]

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

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 (19). We use a pessimistic version of Theorem 4.2: (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 $N_r(\mathcal{P}) \leq k\,\lceil T/\Delta_r \rceil \leq k\,(1+T\frac{\sigma}{r})$. $\square$

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 6.1 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 (17) 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.[7] A clean example is

$$\mathcal{D}((t,y),\,(t',y')) = \min(1,\ \sigma\,|t-t'| + \mathcal{D}_{\mathrm{Y}}(y,y')), \tag{20}$$

where $(Y, \mathcal{D}_{\mathrm{Y}})$ is the arms space. For this example, we can obtain an analog of Corollary 6.1, where the regret bound depends on the covering dimension of the arms space $(Y, \mathcal{D}_{\mathrm{Y}})$.

### 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).

**Uniform marginals.** First we address the general version that we call *drifting MAB with uniform marginals*. Formally, we assume that expected payoffs $\mu(y,\cdot)$ of each arm $y$ evolve over time according to some stochastic process $\Gamma_y$ that satisfies (17). We assume that the processes $\Gamma_y$, $y \in Y$ are mutually independent, and moreover that the marginal distributions $\mu(y,t)$ are uniform on $[0,1]$, for each time $t$ and each arm $y$.[8] We are interested in $\mathbb{E}_\Gamma[\hat{R}(T)]$, average dynamic regret in expectation over the processes $\Gamma_y$.

We obtain a stronger version of (19) via Theorem 4.7. To use this theorem, we need to bound the adjusted $r$-zooming number, call it $N_r$. We show that

$$\mathbb{E}_\Gamma[N_r] = O(kr)\lceil\tfrac{T\sigma}{r}\rceil \text{ and } \left(r < \sigma^{1/3} \Rightarrow N_r = 0\right). \tag{21}$$

Then we obtain a different bound on dynamic regret, which is stronger than Corollary 6.1 for $k < \sigma^{-1/2}$.

**Corollary 6.2.** *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 $T \geq T_0 \geq \sigma^{-2/3}\log\frac{1}{\sigma}$.*

The crux of the proof is to show (21). Interestingly, it involves using all three optimizations in Theorem 4.7: $N_r(\mathcal{P}_{\mu,r})$, $N_r(\mathcal{P}_{\mu,r}\setminus\mathcal{W}_{\mu,r})$ and $N_r^{\mathrm{adj}}(\cdot)$, whereas any two of them do not seem to suffice. The rest is a straightforward computation similar to the one in Corollary 6.1.

**Dynamic MAB.** Let us consider the Dynamic MAB problem from Slivkins and Upfal (2008). Here for each arm $y$ the stochastic process $\Gamma_y$ 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.[9]

[6] This choice of $r_0$ minimizes the inf expression in (5) up to constant factors by equating the two summands.

[7] 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.

[8] E.g. this assumption is satisfied by any Markov Chain on $[0,1]$ with stationary initial distribution.

[9] Slivkins and Upfal (2008) has a slightly more general setup which does not require $1/\sigma_y \in \mathbb{N}$.

According to a well-known fact about random walks that

$$\Pr\left[|\mu(t,y)-\mu(t',y)| \leq O(\sigma_y\,|t-t'|^{1/2}\log T_0)\right] \geq 1-T_0^{-3} \quad \text{if } |t-t'| \leq T_0. \tag{22}$$

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 (22). Namely, we set

$$\mathcal{D}_{T_0}((t,y),\,(t',y')) = \min(1,\; \sigma_y\,|t-t'|^{1/2}\log T_0 + \mathbf{1}_{\{y\neq y'\}}). \tag{23}$$

The following corollary is proved using the same technique as Corollary 6.2:

**Corollary 6.3.** *Consider the Dynamic MAB problem with $k$ arms and volatilities $\sigma_y \equiv \sigma$. Let $\mathtt{ALG}_{T_0}$ denote the contextual zooming algorotihm with period $T_0$ which is parameterized by the similarity space $(\mathcal{P}, \mathcal{D}_{T_0})$. Then $\mathtt{ALG}_{T_0}$ 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 Other applications

**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 4.2.

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.

**Bandit learning-to-rank.** Following a preliminary publication of this paper on `arxiv.org`, contextual zooming has been applied in Slivkins et al. (2010) to bandit learning-to-rank. Interestingly, the "contexts" studied in Slivkins et al. (2010) 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. (2010) 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 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.

**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 $\Pi_t$ over payoff functions. Distributions $\Pi_t$ 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

$$y^*(x) \in \operatorname{argmax}_{y\in Y} \textstyle\sum_{t=1}^{T} \mu_t(x,y), \tag{24}$$

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

$$R(T) \triangleq \textstyle\sum_{t=1}^{T} \mu_t(x_t, y_t) - \mu_t^*(x_t), \quad \text{where} \quad \mu_t^*(x) \triangleq \mu_t(x, y^*(x)). \tag{25}$$

Similarity information is given to an algorithm as a pair of metric spaces: a metric space $(X, \mathcal{D}_{\mathrm{X}})$ on contexts (the *context space*) and a metric space $(Y, \mathcal{D}_{\mathrm{Y}})$ on arms (the *arms space*), which form the product similarity space $(X \times Y, \mathcal{D}_{\mathrm{X}} + \mathcal{D}_{\mathrm{Y}})$. We assume that for each round $t$ functions $\mu_t$ and $\mu_t^*$ are Lipschitz on $(X \times Y, \mathcal{D}_{\mathrm{X}} + \mathcal{D}_{\mathrm{Y}})$ and $(X, \mathcal{D}_{\mathrm{X}})$, 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 (24) is attained by some $y \in Y$. Without loss of generality, $\mathtt{diameter}(X, \mathcal{D}_{\mathrm{X}}) \leq 1$.

Formally, a problem instance consists of metric spaces $(X, \mathcal{D}_{\mathrm{X}})$ and $(Y, \mathcal{D}_{\mathrm{Y}})$, 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.

**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.[10] For a more concrete theorem statement we will assume that the convergence time of `Bandit` is at most $T_0(r) \triangleq c_{\mathrm{Y}}\, r^{-(2+d_{\mathrm{Y}})} \log(\frac{1}{r})$ for some constants $c_{\mathrm{Y}}$ and $d_{\mathrm{Y}}$ that are known to the algorithm. In particular, an algorithm in Kleinberg (2004) achieves this guarantee if $d_{\mathrm{Y}}$ is the $c$-covering dimension of the arms space and $c_{\mathrm{Y}} = O(c^{2+d_{\mathrm{Y}}})$.

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 $\Pi_t$ 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 $d_{\mathrm{Y}} = 0$ and $c_{\mathrm{Y}} = \mathrm{poly}(d)$ (Dani et al., 2007, Abernethy et al., 2008). Likewise, for convex payoffs there exist algorithms with $d_{\mathrm{Y}} = 2$ and $c_{\mathrm{Y}} = O(d)$ (Flaxman et al., 2005). For a bounded number of arms, algorithm EXP3 (Auer et al., 2002b) achieves $d_{\mathrm{Y}} = 0$ and $c_Y = O(\sqrt{|Y|})$.

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

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, in the full version we use a more refined notion which allows to disregard a limited number of "outliers" in $x_{(1..T)}$. Our result is stated as follows:

**Theorem 7.1.** *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*

$$R(T) \leq O(c_{\mathrm{DBL}}^2\, (c_X\, c_Y)^{1/(2+d_X+d_Y)})\ T^{1-1/(2+d_X+d_Y)}(\log T), \tag{26}$$

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

**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 $\mathtt{ALG}_B$ of `Bandit` is created, whose set of "arms" is $Y$. $\mathtt{ALG}_B$ 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 $\mathtt{ALG}_B$ to select an arm $y \in Y$ to be played, and reports the payoff $\pi_t$ back to $\mathtt{ALG}_B$. 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 $B'$.

The analysis of this algorithm (which proves Theorem 7.1) is deferred to the full version.

**Acknowledgements.** 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 have been tremendously useful.

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

## References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization. In *21th COLT*, 2008.

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

P. 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.

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

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

P. Auer, R. Ortner, and C. Szepesvári. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *20th COLT*, 2007.

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

S. Bubeck and R. Munos. Open Loop Optimistic Planning. In *23rd COLT*, 2010.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. Online Optimization in X-Armed Bandits. In *NIPS*, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual Bandits with Linear Payoff Functions. In *14th AISTATS*, 2011.

V. Dani, T. P. Hayes, and S. Kakade. The Price of Bandit Information for Online Optimization. In *NIPS*, 2007.

A. Flaxman, A. Kalai, and H. B. McMahan. Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient. In *16th ACM-SIAM SODA*, 2005.

A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low–distortion embeddings. In *44th IEEE FOCS*, pages 534–543, 2003.

E. Hazan and S. Kale. Better algorithms for benign bandits. In *20th ACM-SIAM SODA*, 2009.

E. Hazan and N. Megiddo. Online Learning with Prior Information. In *20th COLT*, 2007.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th NIPS*, 2004.

R. Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, MIT, Boston, MA, 2005.

R. Kleinberg and A. Slivkins. Sharp Dichotomies for Regret Minimization in Metric Spaces. In *ACM-SIAM SODA*, 2010.

R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In *21st COLT*, 2008a.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-Armed Bandits in Metric Spaces. In *ACM STOC*, 2008b.

L. Kocsis and C. Szepesvari. Bandit Based Monte-Carlo Planning. In *17th ECML*, 2006.

T. Lai and H. Robbins. Asymptotically efficient Adaptive Allocation Rules. *Adv. in Appl. Math.*, 6:4–22, 1985.

J. Langford and T. Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *21st NIPS*, 2007.

A. Lazaric and R. Munos. Hybrid Stochastic-Adversarial On-line Learning. In *22nd COLT*, 2009.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th WWW*, 2010.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *4th WSDM*, 2011.

T. Lu, D. Pál, and M. Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. In *14th AISTATS*, 2010.

O.-A. Maillard and R. Munos. Online Learning in Adversarial Lipschitz Environments. In *ECML PKDD*, 2010.

H. B. McMahan and M. Streeter. Tighter Bounds for Multi-Armed Bandits with Expert Advice. In *22nd COLT*, 2009.

R. Munos and P.-A. Coquelin. Bandit algorithms for tree search. In *23rd UAI*, 2007.

S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for Taxonomies: A Model-based Approach. In *SDM*, 2007.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *25th ICML*, 2008.

P. Rigollet and A. Zeevi. Nonparametric Bandits with Covariates. In *23rd COLT*, 2010.

H. Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

A. Slivkins and E. Upfal. Adapting to a Changing Environment: the Brownian Restless Bandits. In *21st COLT*, 2008.

A. Slivkins, F. Radlinski, and S. Gollapudi. Learning optimally diverse rankings over large document collections. In *27th ICML*, 2010.

C.-C. Wang, S. R. Kulkarni, and H. V. Poor. Bandit problems with side observations. *IEEE Trans. on Automatic Control*, 50(3):338–355, 2005.

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for Infinitely Many-Armed Bandits. In *NIPS*, 2008.

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