Project number IST-25582

CGL

Computational Geometric Learning

Continuum armed bandit problem of few variables in high dimensions

STREP

Information Society Technologies

Period covered: November 1, 2012–October 31, 2013
Date of preparation: November 11, 2013
Date of revision: November 11, 2013
Start date of project: November 1, 2010
Duration: 3 years
Project coordinator name: Joachim Giesen (FSU)
Project coordinator organisation: Friedrich-Schiller-Universität Jena
Jena, Germany

Continuum armed bandit problem of few variables in high dimensions

Hemant Tyagi and Bernd Gärtner

Institute of Theoretical Computer Science,
ETH Zürich (ETHZ), CH-8092 Zürich, Switzerland,
{htyagi,gaertner}@inf.ethz.ch

Abstract. We consider the stochastic and adversarial settings of continuum armed bandits where the arms are indexed by $[0,1]^d$. The reward functions $r : [0,1]^d \to \mathbb{R}$ are assumed to *intrinsically* depend on at most k coordinate variables implying $r(x_1,\ldots,x_d) = g(x_{i_1},\ldots,x_{i_k})$ for distinct and unknown $i_1,\ldots,i_k \in \{1,\ldots,d\}$ and some locally Hölder continuous $g : [0,1]^k \to \mathbb{R}$ with exponent $\alpha \in (0,1]$. Firstly, assuming $(i_1,\ldots,i_k)$ to be fixed across time, we propose a simple modification of the CAB1 algorithm where we construct the discrete set of sampling points to obtain a bound of $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}C(k,d))$ on the regret, with $C(k,d)$ depending at most polynomially in k and sub-logarithmically in d. The construction is based on creating partitions of $\{1,\ldots,d\}$ into k disjoint subsets and is probabilistic, hence our result holds with high probability. Secondly we extend our results to also handle the more general case where $(i_1,\ldots,i_k)$ can change over time and derive regret bounds for the same.

Keywords: Bandit problems, continuum armed bandits, functions of few variables, online optimization.

1 Introduction

In online decision making problems, a player is required to play a strategy, chosen from a given set of strategies S, over a period of n trials or rounds. Each strategy has a reward associated with it specified by a reward function $r : S \to \mathbb{R}$ which typically changes across time in a manner unknown to the player. The aim of the player is to choose the strategies in a manner so as to minimize the *regret* defined as the difference between the total expected reward of the best fixed strategy (not varying with time) and the expected reward of the sequence of strategies played by the player. If the regret after n rounds is sub-linear in n, this implies as $n \to \infty$ that the per-round expected reward of the player asymptotically approaches that of the best strategy. There are many applications of online decision making problems such as routing [1, 2], wireless networks [3], online auction mechanisms [4, 5], statistics (sequential design of experiments [6]) and economics (pricing [7]), to name a few. An important type of online decision making problem is the *multi-armed bandit* problem, where the player only receives the reward associated with the strategy that was played in the round[1]. These problems have been studied extensively when the strategy set S is finite and optimal regret bounds are known within a constant factor [8, 9, 6]. On the other hand, the setting in which S is infinite has been an area of recent attention due to its practical significance. Such problems are referred to as continuum armed bandit problems and are the focus of this paper. Usually S is considered to be a compact subset of a metric space such as $\mathbb{R}^d$. Some applications

[1] Another type is the *best expert problem*, where the entire reward function is revealed to the player at the end of each round.

of these problems are in: (i) online auction mechanism design [4, 5] where the set of feasible prices is representable as an interval and, (ii) online oblivious routing [2] where S is a flow polytope.

For a d-dimensional strategy space it is well known that any multi-armed bandit algorithm will incur worst-case regret of $\Omega(2^d)$ (see [10]). To circumvent this curse of dimensionality, additional assumptions are made on the structure of the reward functions such as linearity (see for example [11]) or convexity (see for example [10]). For these classes of reward functions the regret is typically polynomial in d and sub-linear in n. We consider the setting where the reward function $r : [0,1]^d \to \mathbb{R}$ depends on an unknown subset of k *active* coordinate variables implying $r(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_k})$. The environment is allowed to sample the underlying function g either in an i.i.d manner from some fixed underlying distribution (stochastic) or arbitrarily (adversarial). To the best of our knowledge, such a structure for reward functions has not been considered in the bandit setting previously. On the other hand, there has been significant effort in other fields to develop tractable algorithms for approximating such types of functions (cf. [12, 13] and references within). Our contribution is therefore to combine ideas from different communities and apply them to the bandit setting.

The continuum armed bandit problem was first introduced in [14] for $d = 1$ in the stochastic setting where a regret bound of $o(n^{(2\alpha+1)/(3\alpha+1)+\eta})$ for any $\eta > 0$ was shown for Hölder continuous[2] reward functions with exponent $\alpha \in (0, 1]$. In [5] a lower bound of $\Omega(n^{1/2})$ was proven for this problem. This was then improved upon in [10] where upper and lower bounds of $O(n^{\frac{\alpha+1}{2\alpha+1}}(\log n)^{\frac{\alpha}{2\alpha+1}})$ and $\Omega(n^{\frac{\alpha+1}{2\alpha+1}})$ were derived for both stochastic and adversarial settings. [15] considered a class of reward functions with additional smoothness properties and derived a regret bound of $O(n^{1/2})$ which is also optimal. In [16] the case $d = 1$ was treated, with the reward function assumed to only satisfy a local Hölder condition around the maximum $\mathbf{x}^*$ with exponent $\alpha \in (0, \infty)$. Under these assumptions the authors considered a modification of Kleinberg's CAB1 algorithm [10] and achieved a regret bound of $O(n^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}}(\log n)^{\frac{\alpha}{1+2\alpha-\alpha\beta}})$ for some known $0 < \beta < 1$. In [17, 18] the authors studied a very general setting in which S forms a metric space, with the reward function assumed to satisfy a Lipschitz condition with respect to this metric and derived close to optimal regret bounds.

Our contributions are twofold. Firstly, we prove that when $(i_1, \ldots, i_k)$ is fixed across time but unknown to the player, then a simple modification of the `CAB1` algorithm can be used to achieve a regret bound [3] of $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}C(k,d))$ where $\alpha \in (0, 1]$ denotes the exponent of Hölder continuity of the reward functions. The factor $C(k,d) = O(\text{poly}(k) * o(\log d))$ captures the uncertainty of not knowing the k active coordinates. The modification is in the manner of discretization of $[0,1]^d$ for which we consider a probabilistic construction based on creating *partitions* of $\{1, \ldots, d\}$ into k disjoint subsets. The above bound holds for both the stochastic (underlying g is sampled in an i.i.d manner) and the adversarial (underlying g chosen arbitrarily at each round) models. Secondly, we extend our results to handle the more general setting where an adversary chooses some sequence of k-tuples $(\mathbf{i}_t)_{t=1}^n = (i_{1,t}, \ldots, i_{k,t})_{t=1}^n$ before the start of plays. For this setting we derive a regret bound of $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}H[(\mathbf{i}_t)_{t=1}^n]C(k,d))$ where $H[(\mathbf{i}_t)_{t=1}^n]$ denotes the "hardness" [4] of the sequence $(\mathbf{i}_t)_{t=1}^n$. Furthermore, in case $H[(\mathbf{i}_t)_{t=1}^n] \leq S$ for some $S > 0$ known to player, the regret bound then improves to $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}S^{\frac{\alpha}{2\alpha+k}}C(k,d))$.

[2] A function $r : S \to \mathbb{R}$ is Hölder continuous if $|r(\mathbf{x}) - r(\mathbf{y})| \leq L \parallel \mathbf{x} - \mathbf{y} \parallel^\alpha$ for constants $L > 0$, $\alpha \in (0, 1]$ and any $\mathbf{x}, \mathbf{y} \in S$.

[3] See Remark 1 in Section 3 for discussion on how the $\log n$ term can be removed.

[4] See Definition 3 in Section 4.

The rest of the paper is organized as follows. In Section 2 we define the problem statement formally and outline our main results. In Section 3 we present an analysis for the setting when the active k coordinates are fixed across time, including the construction of the discrete strategy sets. In Section 4 we consider the setting where the active k coordinates change across time. Finally in Section 5 we summarize our results and provide directions for future work.

2 Problem Setup and Main Results

The compact set of strategies $S = [0,1]^d \subset \mathbb{R}^d$ is available to the player. At each time step $t = 1, \ldots, n$, a reward function $r_t : S \to \mathbb{R}$ is chosen by the environment. Upon playing a strategy $\mathbf{x}_t \in [0,1]^d$, the player receives the reward $r_t(\mathbf{x}_t)$ at time step t. For some $k \leq d$, we assume each r_t to depend on a fixed but unknown subset of k variables implying $r_t(x_1, \ldots, x_d) = g_t(x_{i_1}, \ldots, x_{i_k})$ where $(i_1, \ldots, i_k)$ is a k-tuple with distinct integers $i_j \in \{1, \ldots, d\}$ and $g_t : [0,1]^k \to \mathbb{R}$. For simplicity of notation, we denote the set of such k-tuples of the set $\{1, \ldots, d\}$ by $\mathcal{T}_k^d$ and the ℓ_2 norm by $\| \cdot \|$. We assume that k is known to the player, however it suffices to know a bound for k as well. The second assumption that we make is on the smoothness property of the reward functions.

Definition 1. *A function $f : [0,1]^k \to \mathbb{R}$ is locally uniformly Hölder continuous with constant $0 \leq L < \infty$, exponent $0 < \alpha \leq 1$, and restriction $\delta > 0$ if we have for all $\mathbf{u}, \mathbf{u}' \in [0,1]^k$ with $\| \mathbf{u} - \mathbf{u}' \| \leq \delta$ that $|f(\mathbf{u}) - f(\mathbf{u}')| \leq L \| \mathbf{u} - \mathbf{u}' \|^\alpha$. We denote the class of such functions f as $\mathcal{C}(\alpha, L, \delta, k)$.*

The function class defined in Definition 1 was also considered in [14, 10] and is a generalization of Lipschitz continuity (obtained for $\alpha = 1$). We now define the two models that we analyze in this paper. These models describe how the reward functions g_t are generated at each time step t.

Stochastic model. The reward functions g_t are considered to be i.i.d samples from some fixed but unknown probability distribution over functions $g : [0,1]^k \to \mathbb{R}$. We define the expectation of the reward function as $\bar{g}(\mathbf{u}) = \mathbb{E}[g(\mathbf{u})]$ where $\mathbf{u} \in [0,1]^k$. We require $\bar{g}$ to belong to $\mathcal{C}(\alpha, L, \delta, k)$ and note that the individual samples g_t need not necessarily be Hölder continuous. The optimal strategy $\mathbf{x}^*$ is then defined as follows.

$$\mathbf{x}^* := \text{argmax}_{\mathbf{x} \in [0,1]^d} \mathbb{E}[r(\mathbf{x})] = \text{argmax}_{\mathbf{x} \in [0,1]^d} \bar{g}(x_{i_1}, \ldots, x_{i_k}). \tag{1}$$

Adversarial model. The reward functions $g_t : [0,1]^k \to [0,1]$ are a fixed sequence of functions in $\mathcal{C}(\alpha, L, \delta, k)$ chosen arbitrarily by an *oblivious* adversary i.e., an adversary oblivious to the actions of the player. The optimal strategy $\mathbf{x}^*$ is then defined as follows.

$$\mathbf{x}^* := \text{argmax}_{\mathbf{x} \in [0,1]^d} \sum_{t=1}^{n} r_t(\mathbf{x}) = \text{argmax}_{\mathbf{x} \in [0,1]^d} \sum_{t=1}^{n} g_t(x_{i_1}, \ldots, x_{i_k}). \tag{2}$$

Given the above models we measure the performance of a player over n rounds in terms of the *regret* defined as

$$R(n) := \sum_{t=1}^{n} \mathbb{E}\left[r_t(\mathbf{x}^*) - r_t(\mathbf{x}_t)\right] = \sum_{t=1}^{n} \mathbb{E}\left[g_t(\mathbf{x}^*_{i_1}, \ldots, \mathbf{x}^*_{i_k}) - g_t(\mathbf{x}^{(t)}_{i_1}, \ldots, \mathbf{x}^{(t)}_{i_k})\right]. \tag{3}$$

In (3) the expectation is defined over the random choices of g_t for the stochastic model and the random choice of the strategy $\mathbf{x}_t$ at each time t by the player.

Main results. The main results of our work are as follows. Firstly, assuming that the k-tuple $(i_1, \ldots, i_k) \in \mathcal{T}_k^d$ is chosen once at the beginning of play and kept fixed thereafter, we provide in the form of Theorem 1 a bound on the regret which is $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}C(k,d))$ where $C(k,d) = O(\text{poly}(k) * o(\log d))$. This bound holds for both the stochastic and adversarial models and is *almost optimal*. To see this, we note that [19] showed a precise exponential lower bound of $\Omega(n^{\frac{d+1}{d+2}})$ after $n = \Omega(2^d)$ plays for stochastic continuum armed bandits with d-variate Lipschitz continuous reward functions defined over $[0,1]^d$. In our setting though, the reward functions depend on an unknown subset of k coordinate variables hence any algorithm after $n = \Omega(2^k)$ plays would incur worst case regret of $\Omega(n^{\frac{k+1}{k+2}})$ which is still mild if $k \ll d$. We see that our upper bound matches this lower bound for the case of Lipschitz continuous reward functions ($\alpha = 1$) up to a mild factor of $(\log n)^{\frac{1}{2+k}}C(k,d)$. We also note that the $o(\log d)$ factor in (4) accounts for the uncertainty in not knowing which k coordinates are active from $\{1, \ldots, d\}$.

Theorem 1. *Given that the k-tuple $(i_1, \ldots, i_k) \in \mathcal{T}_k^d$ is kept fixed across time but unknown to the player, the algorithm `CAB(d,k)` incurs a regret of*

$$O\left(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}k^{\frac{\alpha(k+6)}{2(2\alpha+k)}}e^{\frac{k\alpha}{2\alpha+k}}(\log d)^{\frac{\alpha}{2\alpha+k}}\right) \tag{4}$$

after n rounds of play with high probability for both the stochastic and adversarial models.

The above result is proven in Section 3 along with a description of the `CAB(d,k)` algorithm which achieves this bound. The main idea here is to discretize $[0,1]^d$ by first constructing a family of partitions $\mathcal{A}$ of $\{1, \ldots, d\}$ with each partition consisting of k disjoint subsets. The construction is probabilistic and the resulting $\mathcal{A}$ satisfies an important property (with high probability) namely the *Partition Assumption* as described in Section 3. In particular we have that $|\mathcal{A}|$ is $O(ke^k \log d)$ resulting in a total of $M^k|\mathcal{A}|$ sampling points for some integer $M > 0$. This discrete strategy set is then used with a finite armed bandit algorithm such as UCB-1 [9] for the stochastic setting and Exp3 [8] for the adversarial setting, to achieve the regret bound of Theorem 1.

Secondly we extend our results to the setting where $(i_1, \ldots, i_k)$ can change over time. Considering that an oblivious adversary chooses arbitrarily before the start of plays a sequence of k tuples $(\mathbf{i}_t)_{t=1}^n = (i_{1,t}, \ldots, i_{k,t})_{t=1}^n$ of *hardness* $H[(\mathbf{i}_t)_{t=1}^n] \leq S$ (see Definition 3 in Section 4) with $S > 0$ known to the player, we show how Algorithm `CAB(d,k)` can be adapted to this setting to achieve a regret bound of $O\left(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}S^{\frac{\alpha}{2\alpha+k}}C(k,d)\right)$. Hardness of a sequence is defined as the number of adjacent elements with different values. In case the player has no knowledge of S, the regret bound then changes to $O(n^{\frac{\alpha+k}{2\alpha+k}}(\log n)^{\frac{\alpha}{2\alpha+k}}H[(\mathbf{i}_t)_{t=1}^n]C(k,d))$. Although our bound becomes trivial when $H[(\mathbf{i}_t)_{t=1}^n]$ is close to n (as one would expect), we can still achieve sub-linear regret when $H[(\mathbf{i}_t)_{t=1}^n]$ is small relative to n. We again consider a discretization of the space $[0,1]^d$ constructed using the family of partitions $\mathcal{A}$ mentioned earlier. The difference lies in now using the Exp3.S algorithm [20] on the discrete strategy set, which in contrast to the Exp3 algorithm is designed to control regret against arbitrary sequences. This is described in Section 4.

3 Analysis when k active coordinates are fixed across time

We begin with the setting where the set of active k coordinates is fixed across time. The very core of our analysis involves the usage of a specific family of partitions $\mathcal{A}$ of $\{1, \ldots, d\}$ where each $\mathbf{A} \in \mathcal{A}$ consists of k disjoint subsets $(A_1, \ldots, A_k)$. In particular

we require $\mathcal{A}$ to satisfy an important property namely the *partition assumption* defined below.

Definition 2. *A family of partitions $\mathcal{A}$ of $\{1,\ldots,d\}$ into k disjoint subsets is said to satisfy the partition assumption if for any k distinct integers $i_1, i_2, \ldots, i_k \in \{1,\ldots,d\}$, there exists a partition $\mathbf{A} = (A_1,\ldots,A_k)$ in $\mathcal{A}$ such that each set in $\mathbf{A}$ contains exactly one of $i_1, i_2, \ldots, i_k$.*

The above definition is known as *perfect hashing* in theoretical computer science and is widely used such as in finding juntas [21]. There exists a fairly simple probabilistic method using which one can construct $\mathcal{A}$ consisting of $O(ke^k \log d)$ partitions satisfying the partition assumption property with high probability (see for example[5], Section 5 in [12]). For our purposes, we consider the aforementioned probabilistic construction. However, there also exist deterministic constructions resulting in larger family sizes such as the one proposed in [22] where a family of size $O(k^{O(\log k)} e^k \log d)$ is constructed deterministically in time $poly(d,k)$. For details we refer the reader to the full version of this paper [23].

Constructing strategy set $\mathcal{P}_M$ using $\mathcal{A}$. Suppose we are given a family of partitions $\mathcal{A}$ satisfying the partition assumption. Then using $\mathcal{A}$ we construct the discrete set of strategies $\mathcal{P}_M \in [0,1]^d$ for some fixed integer $M > 0$ as follows.

$$\mathcal{P}_M := \left\{ \frac{1}{M} \sum_{j=1}^{k} \alpha_j \chi_{\mathbf{A}_j} ; \alpha_j \in \{1,\ldots,M\}, (\mathbf{A}_1,\ldots,\mathbf{A}_k) \in \mathcal{A} \right\} \subset [0,1]^d \qquad (5)$$

Note that a strategy $\mathbf{x} = \frac{1}{M}\sum_{j=1}^{k} \alpha_j \chi_{\mathbf{A}_j}$ has coordinate value $\frac{1}{M}\alpha_j$ at each of the coordinate indices in A_j. Therefore we see that for each partition $\mathbf{A} \in \mathcal{A}$ we have M^k strategies implying a total of $M^k|\mathcal{A}|$ strategies in $\mathcal{P}_M$.

Algorithm 1 Algorithm CAB(d,k)

$T = 1$
Construct family of partitions $\mathcal{A}$ satisfying partition assumption
while $T \leq n$ **do**
  $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (\log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}} \right)^{2/(2\alpha+k)} \right\rceil$
  – Create $\mathcal{P}_M$ using $\mathcal{A}$
  – Initialize **MAB** with $\mathcal{P}_M$
  for $t = T, \ldots, \min(2T-1, n)$ **do**
    – get $\mathbf{x}_t$ from **MAB**
    – Play $\mathbf{x}_t$ and get $r_t(\mathbf{x}_t)$
    – Feed $r_t(\mathbf{x}_t)$ back to **MAB**
  end for
  $T = 2T$
end while

Projection property. An important property of the strategy set $\mathcal{P}_M$ is the following. Given any k-tuple of distinct indices $(i_1,\ldots,i_k)$ with $i_j \in \{1,\ldots,d\}$ and any integers $1 \leq n_1,\ldots,n_k \leq M$, there is a strategy $\mathbf{x} \in \mathcal{P}_M$ such that

$$(x_{i_1},\ldots,x_{i_k}) = \left(\frac{n_1}{M},\ldots,\frac{n_k}{M}\right).$$

[5] In [12] the authors consider the significantly different *function approximation* problem as opposed to our setting of online optimization.

To see this, one can simply take a partition $\mathbf{A} = (A_1, \ldots, A_k)$ from $\mathcal{A}$ such that each i_j is in a different set A_j for $j = 1, \ldots, k$. Then setting appropriate $\alpha_j = n_j$ when $i_j \in A_j$ we get that coordinate i_j of $\mathbf{x}$ has the value n_j/M.

Upper bound on regret. We now describe our Algorithm `CAB(d,k)` and provide bounds on its regret. Note that the outer loop is a standard doubling trick which is used as the player has no knowledge of the time horizon n. Observe that before the start of the inner loop of duration T, the player constructs the finite strategy set $\mathcal{P}_M$, where M increases progressively with T. Within the inner loop, the problem reduces to a finite armed bandit problem. The **MAB** routine can be any standard multi-armed bandit algorithm such as UCB-1 (stochastic model) or Exp3 (adversarial model). The main idea is that for increasing values of M, we would have for any $\mathbf{x}^*$ and any $(i_1, \ldots, i_k)$ the existence of an arbitrarily close point to $(x^*_{i_1}, \ldots, x^*_{i_k})$ in $\mathcal{P}_M$. This follows from the projection property of $\mathcal{P}_M$. Coupled with the Hölder continuity of the reward functions this then ensures that the **MAB** routine progressively plays strategies closer and closer to $\mathbf{x}^*$ leading to a bound on regret. The algorithm is motivated by the `CAB1` algorithm [10], however unlike the equi-spaced sampling done in `CAB1` we consider a probabilistic construction of the discrete set of sampling points based on partitions of $\{1, \ldots, d\}$. Now for the stochastic setting, we make the following assumption on the distribution from which the random samples g are generated.

Assumption 1 *We assume that there exist constants $\zeta, s_0 > 0$ so that*

$$\mathbb{E}[e^{s(g(\mathbf{u})-\bar{g}(\mathbf{u}))}] \leq e^{\frac{1}{2}\zeta^2 s^2} \quad \forall s \in [-s_0, s_0], \mathbf{u} \in [0,1]^k.$$

The above assumption was considered in [10] for the case $d = 1$ and allows us to consider reward functions g_t whose range is not bounded. Note that the mean reward $\bar{g}$ is assumed to be Hölder continuous and is therefore bounded. We now present in the following lemma the regret bound incurred within an inner loop of duration T.

Lemma 1. *Given that $(i_1, \ldots, i_k)$ is fixed across time then if the strategy set $\mathcal{P}_M$ is used with (i) the UCB-1 algorithm for the stochastic setting or, (ii) the Exp3 algorithm for the adversarial setting, we have for the choice $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (\log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}} \right)^{\frac{2}{2\alpha+k}} \right\rceil$ that the regret incurred by the player after T rounds is given by $R(T) = O\left(T^{\frac{\alpha+k}{2\alpha+k}} (\log T)^{\frac{\alpha}{2\alpha+k}} k^{\frac{\alpha(k+6)}{2(2\alpha+k)}} e^{\frac{k\alpha}{2\alpha+k}} (\log d)^{\frac{\alpha}{2}} \right.$*

Proof. For some $\mathbf{x}' \in \mathcal{P}_M$ we can split $R(T)$ into $R_1(T) + R_2(T)$ where:

$$R_1(T) = \sum_{t=1}^{T} \mathbb{E}[g_t(x^*_{i_1}, \ldots, x^*_{i_k}) - g_t(x'_{i_1}, \ldots, x'_{i_k})], \tag{6}$$

$$R_2(T) = \sum_{t=1}^{T} \mathbb{E}[g_t(x'_{i_1}, \ldots, x'_{i_k}) - g_t(x^{(t)}_{i_1}, \ldots, x^{(t)}_{i_k})]. \tag{7}$$

For the k tuple $(i_1, \ldots, i_k) \in \mathcal{T}^d_k$, there exists $\mathbf{x}' \in \mathcal{P}_M$ with $x'_{i_1} = \frac{\alpha_1}{M}, \ldots, x'_{i_k} = \frac{\alpha_k}{M}$ where $\alpha_1, \ldots, \alpha_k$ are such that $|\alpha_j/M - x^*_{i_j}| < (1/M)$. This follows from the projection property of $\mathcal{A}$. On account of the Hölder continuity of reward functions we then have that

$$\mathbb{E}[g_t(x^*_{i_1}, \ldots, x^*_{i_k}) - g_t(x'_{i_1}, \ldots, x'_{i_k})] < L \left(\left(\frac{1}{M} \right)^2 k \right)^{\alpha/2}.$$

In other words, $R_1(T) = O(Tk^{\alpha/2}M^{-\alpha})$. In order to bound $R_2(T)$, we note that the problem has reduced to a $|\mathcal{P}_M|$-armed bandit problem. Specifically we note from (7) that we are comparing against a suboptimal strategy $\mathbf{x}'$ instead of the optimal one in $\mathcal{P}_M$. Hence $R_2(T)$ can be bounded by using existing bounds for finite-armed bandit problems. Now for the stochastic setting we can employ the UCB-1 algorithm [9] and play at each t a strategy $\mathbf{x}_t \in \mathcal{P}_M$. In particular, on account of Assumption 1, it can be shown that $R_2(T) = O(\sqrt{|\mathcal{P}_M|T\log T})$ (Theorem 3.1, [10]). For the adversarial setting we can employ the Exp3 algorithm [8] so that $R_2(T) = O(\sqrt{|\mathcal{P}_M|T\log|\mathcal{P}_M|})$. Combining the bounds for $R_1(T)$ and $R_2(T)$ and recalling that $|\mathcal{P}_M| = O(M^k k e^k \log d)$ we obtain:

$$R(T) = O(TM^{-\alpha}k^{\alpha/2} + \sqrt{M^k k e^k \log d\, T\log T}) \text{ (stochastic) and,} \tag{8}$$

$$R(T) = O(TM^{-\alpha}k^{\alpha/2} + \sqrt{M^k k e^k \log d\, T\log(M^k k e^k \log d)}). \text{ (adversarial)} \tag{9}$$

Plugging $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (\log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}}\right)^{\frac{2}{2\alpha+k}} \right\rceil$ in (8) and (9) we obtain the stated bound on $R(T)$ for the respective models. □

Lastly equipped with the above bound we have that the regret incurred by Algorithm 1 over n plays is given by:

$$\sum_{i=0,T=2^i}^{i=\log n} R(T) = O\left(n^{\frac{\alpha+k}{2\alpha+k}} (\log n)^{\frac{\alpha}{2\alpha+k}} k^{\frac{\alpha(k+6)}{2(2\alpha+k)}} e^{\frac{k\alpha}{2\alpha+k}} (\log d)^{\frac{\alpha}{2\alpha+k}}\right).$$

Remark 1. For the adversarial setting we can use the INF algorithm of [24] as the MAB routine in our algorithm and get rid of the $\log n$ factor from the regret bound. The same holds for the stochastic setting, if the range of the reward functions was restricted to be $[0, 1]$. When the range of the reward functions is $\mathbb{R}$, as is the case in our setting, it seems possible to consider a variant of the MOSS algorithm [24] along with Assumption 2 on the distribution of the reward functions (using proof techniques similar to [25]), to remove the $\log n$ factor from the regret bound.

4 Analysis when k active coordinates change across time

We now consider a more general *adversarial* setting where the k tuple $(i_1, \ldots, i_k)$ is allowed to change over time. Formally this means that the reward functions $(r_t)_{t=1}^n$ now have the form $r_t(x_1, \ldots, x_d) = g_t(x_{i_{1,t}}, \ldots, x_{i_{k,t}})$ where $(i_{1,t}, \ldots, i_{k,t})_{t=1}^n$ denotes the sequence of k-tuples chosen by the adversary before the start of plays. However we assume that this sequence of k-tuples is not "hard" meaning that it contains a small number of consecutive pairs (relative to the number of rounds n) with different values. Furthermore, $r_t : [0, 1]^d \to [0, 1]$ with $g_t : [0, 1]^k \to [0, 1]$ where $g_t \in \mathcal{C}(\alpha, \delta, L, k)$. We now formally present the definition of hardness of a sequence.

Definition 3. *For any set $\mathcal{B}$ we define the hardness of the sequence $(b_1, \ldots, b_n) \in \mathcal{B}^n$ by:*

$$H[b_1, \ldots, b_n] := 1 + |\{1 \le l < n : b_l \ne b_{l+1}\}|. \tag{10}$$

The above definition is borrowed from Section 8 in [20] where the authors considered the non-stochastic multi armed bandit problem, and employed the definition to characterize the hardness of a sequence of actions against which the regret of the players actions is measured. In our setting, we consider the sequence of k-tuples chosen by the adversary to be at most S-hard, meaning that $H[(i_{1,t}, \ldots, i_{k,t})_{t=1}^n] \le S$ for some $S > 0$, and also assume that S is known to the player. We now proceed to show how a slight modification of Algorithm CAB(d, k) can be used to

derive a bound on the regret in this setting. Recall that the optimal strategy $\mathbf{x}^* := \text{argmax}_{\mathbf{x}\in[0,1]^d} \sum_{t=1}^n g_t(x_{i_{1,t}}, \ldots, x_{i_{k,t}})$. Since the sequence of k-tuples is S-hard, this in turn implies for any $\mathbf{x}^*$ that $H[(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})_{t=1}^n] \leq S$. Therefore we can now consider this as a setting where the players regret is measured against an S-hard sequence $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})_{t=1}^n$.

Now the player does not know which k-tuple is chosen at each time t. Hence we again construct the discrete strategy set $\mathcal{P}_M$ (as defined in (5)) using the family of partitions $\mathcal{A}$ of $\{1, \ldots, d\}$. By construction, we will have for any $\mathbf{x} \in [0,1]^d$ and any k-tuple $(i_1, \ldots, i_k)$, the existence of a point $\mathbf{z}$ in $\mathcal{P}_M$ such that $(z_{i_1}, \ldots, z_{i_k})$ approximates $(x_{i_1}, \ldots, x_{i_k})$ arbitrarily well for increasing values of M. Hence, for the optimal sequence $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})_{t=1}^n$, we have the existence of a sequence of points $(\mathbf{z}^{(t)})_{t=1}^n$ where $\mathbf{z}^{(t)} \in \mathcal{P}_M$ with the following two properties.

1. *S-hardness.* $H[(z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}})_{t=1}^n] \leq S$. This follows easily from the S-hardness of the sequence $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})_{t=1}^n$ and by choosing for each $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})$ a corresponding $\mathbf{z}^{(t)} \in \mathcal{P}_M$ such that $\parallel (x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}}) - (z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}}) \parallel$ is minimized.
2. *Approximation property.* $\parallel (x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}}) - (z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}}) \parallel = O(k^{\alpha/2} M^{-\alpha})$. This is easily verifiable via the projection property of the set $\mathcal{P}_M$.

Therefore by employing the Exp3.S algorithm [20] on the strategy set $\mathcal{P}_M$ we reduce the problem to a finite armed adversarial bandit problem where the players regret measured against the S-hard sequence $(z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}})_{t=1}^n$ is bounded from above. The approximation property of this sequence (as explained above) coupled with the Hölder continuity of g_t ensures in turn that the players regret against the original sequence $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})_{t=1}^n$ is also bounded. With this in mind we present the following lemma, which formally states a bound on regret after T rounds of play.

Lemma 2. *Given the above setting and assuming that:*

1. *the sequence of k-tuples $(i_{1,t}, \ldots, i_{k,t})_{t=1}^n$ is at most S-hard and,*
2. *the Exp3.S algorithm is used along with the strategy set $\mathcal{P}_M$,*

we have for the choice $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (S \log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}} \right)^{\frac{2}{2\alpha+k}} \right\rceil$ *that the regret incurred by the player after T rounds is given by:*

$$R(T) = O\left(T^{\frac{\alpha+k}{2\alpha+k}} (\log T)^{\frac{\alpha}{2\alpha+k}} k^{\frac{\alpha(k+6)}{2(2\alpha+k)}} e^{\frac{k\alpha}{2\alpha+k}} (S \log d)^{\frac{\alpha}{2\alpha+k}} \right).$$

Proof. At each time t, for some $\mathbf{z}^{(t)} \in \mathcal{P}_M$ we can split $R(T)$ into $R_1(T) + R_2(T)$ where $R_1(T) = \mathbb{E}[\sum_{t=1}^T g_t(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}}) - g_t(z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}})]$ and $R_2(T) = \mathbb{E}[\sum_{t=1}^T g_t(z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}}) - g_t(x^{(t)}_{i_{1,t}}, \ldots, x^{(t)}_{i_{k,t}})]$.

Let us consider $R_1(T)$ first. As before, from the projection property of $\mathcal{A}$ we have for each $(x^*_{i_{1,t}}, \ldots, x^*_{i_{k,t}})$, that there exists $\mathbf{z}^{(t)} \in \mathcal{P}_M$ with $z^{(t)}_{i_{1,t}} = \frac{\alpha^{(t)}_1}{M}, \ldots, z^{(t)}_{i_{k,t}} = \frac{\alpha^{(t)}_k}{M}$ where $\alpha^{(t)}_1, \ldots, \alpha^{(t)}_k$ are such that $|\alpha^{(t)}_j / M - x^*_{i_{j,t}}| < (1/M)$ holds for $j = 1, \ldots, k$ and each $t = 1, \ldots, n$. Therefore from Hölder continuity of g_t we obtain $R_1(T) = O(Tk^{\alpha/2} M^{-\alpha})$. It remains to bound $R_2(T)$. To this end, note that the sequence $(z^{(t)}_{i_{1,t}}, \ldots, z^{(t)}_{i_{k,t}})_{t=1}^n$ with $\mathbf{z}^{(t)} \in \mathcal{P}_M$ is at most S-hard. Hence the problem has reduced to a $|\mathcal{P}_M|$ armed adversarial bandit problem with a S-hard optimal sequence of plays against which the regret of the player is to be bounded. This is accomplished by using the Exp3.S algorithm of [20] which is designed to control regret against *any* S-hard sequence of plays. In particular from Corollary 8.3 of [20] we have that $R_2(T) = O(\sqrt{S|\mathcal{P}_M| T \log(|\mathcal{P}_M| T)})$. Combining the bounds for $R_1(T)$ and $R_2(T)$

and recalling that $|\mathcal{P}_M| = O(M^k k e^k \log d)$ we obtain the following expression for $R(T)$:

$$R(T) = O(Tk^{\alpha/2}M^{-\alpha} + \sqrt{STM^k k e^k \log d \log(TM^k k e^k \log d)}). \tag{11}$$

Lastly after plugging in the value $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (S \log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}}\right)^{\frac{2}{2\alpha+k}} \right\rceil$ in (11), we obtain the stated bound on $R(T)$. □

By employing Algorithm 1 with **MAB** sub-routine being the Exp3.S algorithm, we have that its regret over n plays is given by

$$\sum_{i=0, T=2^i}^{i=\log n} R(T) = O\left(n^{\frac{\alpha+k}{2\alpha+k}} (\log n)^{\frac{\alpha}{2\alpha+k}} k^{\frac{\alpha(k+6)}{2(2\alpha+k)}} e^{\frac{k\alpha}{2\alpha+k}} (S \log d)^{\frac{\alpha}{2\alpha+k}}\right).$$

Remark 2. In case the player does not know S, then a regret of

$$R(n) = O\left(n^{\frac{\alpha+k}{2\alpha+k}} (\log n)^{\frac{\alpha}{2\alpha+k}} k^{\frac{\alpha(k+6)}{2(2\alpha+k)}} e^{\frac{k\alpha}{2\alpha+k}} (\log d)^{\frac{\alpha}{2\alpha+k}} H[(\mathbf{i}_t)_{t=1}^n]\right)$$

would be incurred by Algorithm 1 with the **MAB** routine being the Exp3.S algorithm and for the choice $M = \left\lceil \left(k^{\frac{\alpha-3}{2}} e^{-\frac{k}{2}} (\log d)^{-\frac{1}{2}} \sqrt{\frac{T}{\log T}}\right)^{\frac{2}{2\alpha+k}} \right\rceil$. Here $\mathbf{i}_t$ is shorthand notation for $(i_{1,t}, \ldots, i_{k,t})$. This can be verified easily along the lines of the proof of Lemma 2 by noting that on account of Corollary 8.2 of [20], we have $R_2(T) = O(H[(\mathbf{i}_t)_{t=1}^n] \sqrt{|\mathcal{P}_M| T \log(|\mathcal{P}_M| T)})$.

5 Concluding Remarks

In this work we considered continuum armed bandit problems for the stochastic and adversarial settings where the reward function $r : [0,1]^d \to \mathbb{R}$ depends at each time step on only k out of the d coordinate variables. We proposed an algorithm and proved regret bounds, both for the setting when the the active k coordinates remain fixed across time and also for the more general scenario when they can change over time. There are several interesting lines of future work. Firstly for the case when $(i_1, \ldots, i_k)$ is fixed across time it would be interesting to investigate whether the dependence of regret on k and dimension d achieved by our algorithm, is optimal or not. Secondly, for the case when $(i_1, \ldots, i_k)$ can also change with time, it would be interesting to derive lower bounds on regret to know what the optimal dependence on the hardness of the sequence of k tuples is.

Acknowledgments. The authors thank Sebastian Stich for the helpful discussions and comments on the manuscript and Fabrizio Grandoni for making us aware of deterministic constructions of perfect hash functions. The project CG Learning acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 255827.

References

1. B. Awerbuch and R. Kleinberg. Near-optimal adaptive routing: Shortest paths and geometric generalizations. In *Proceedings of ACM Symposium on Theory of Computing*, 2004.

2. N. Bansal, A. Blum, S. Chawla, and A. Meyerson. Online oblivious routing. In *Proceedings of ACM Symposium in Parallelism in Algorithms and Architectures*, pages 44–49, 2003.
3. C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems*, 2003.
4. A. Blum, V. Kumar, A. Rudra, and F. Wu. Online learning in online auctions. In *Proceedings of 14th Symp. on Discrete Alg.*, pages 202–204, 2003.
5. R. Kleinberg and T. Leighton. The value of knowing a demand curve: bounds on regret for online posted-price auctions. In *Proceedings of Foundations of Computer Science, 2003.*, pages 594–605, 2003.
6. T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocations rules. In *Proceedings of Adv. in Appl. Math.*, volume 6, pages 4–22, 1985.
7. M. Rothschild. A two-armed bandit theory of market pricing. In *Journal of Economic Theory*, volume 9, pages 185–202, 1974.
8. P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of 36th Annual Symposium on Foundations of Computer Science, 1995*, pages 322–331, 1995.
9. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.
10. R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems*, 2004.
11. J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.
12. R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constr. Approx.*, 33:125–143, 2011.
13. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15:1373–1396, 2003.
14. R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.
15. E.W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *Automatic Control, IEEE Transactions on*, 54:1243–1253, 2009.
16. P. Auer, R. Ortner, and C. Szepesvari. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of 20th Conference on Learning Theory (COLT)*, pages 454–468, 2007.
17. R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 681–690, 2008.
18. S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. X-armed bandits. *Journal of Machine Learning Research (JMLR)*, 12:1587–1627, 2011.
19. S. Bubeck, G. Stoltz, and J.Y. Yu. Lipschitz bandits without the Lipschitz constant. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, pages 144–158, 2011.
20. P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.
21. E. Mossel, R. O'Donnell, and R. Servedio. Learning juntas. In *Proceedings of the thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC, pages 206–212. ACM, 2003.
22. M. Naor, L.J. Schulman, and A. Srinivasan. Splitters and near-optimal derandomization. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science, 1995.*, pages 182–191, 1995.
23. H. Tyagi and B. Gärtner. Continuum armed bandit problem of few variables in high dimensions. *CoRR*, abs/1304.5793, 2013.
24. J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.
25. R.D. Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, MIT, Boston, MA, 2005.