Improved Rates for the Stochastic Continuum-Armed Bandit Problem

Peter Auer[1], Ronald Ortner[1], and Csaba Szepesvári[2]

[1] University of Leoben, A-8700 Leoben, Austria
auer@unileoben.ac.at, ronald.ortner@unileoben.ac.at
[2] University of Alberta, Edmonton T6G 2E8, Canada
szepesva@cs.ualberta.ca

Abstract. Considering one-dimensional continuum-armed bandit problems, we propose an improvement of an algorithm of Kleinberg and a new set of conditions which give rise to improved rates. In particular, we introduce a novel assumption that is complementary to the previous smoothness conditions, while at the same time smoothness of the mean payoff function is required only at the maxima. Under these new assumptions new bounds on the expected regret are derived. In particular, we show that apart from logarithmic factors, the expected regret scales with the square-root of the number of trials, provided that the mean payoff function has finitely many maxima and its second derivatives are continuous and non-vanishing at the maxima. This improves a previous result of Cope by weakening the assumptions on the function. We also derive matching lower bounds. To complement the bounds on the expected regret, we provide high probability bounds which exhibit similar scaling.

1 Introduction

We consider continuum-armed bandit problems defined by some unknown distribution-family $P(\cdot|x)$, indexed by $x \in [0,1]$. In each trial $t = 1, 2, \ldots$ the learner chooses $X_t \in [0,1]$ and receives return $Y_t \sim P(\cdot|X_t)$. We assume that Y_t is independent of $\mathcal{F}_{t-1} = \sigma(X_1, Y_1, \ldots, X_{t-1}, Y_{t-1})$ given X_t. Furthermore, the returns are assumed to be uniformly bounded, say $Y_t \in [0,1]$.

The goal of the learner is to maximize her expected return. Let the mean return at x be

$$b(x) \triangleq \mathbb{E}\left[Y_1 \,|\, X_1 = x\right],$$

where we assume that $b : [0,1] \to [0,1]$ is measurable. Let $b^* \triangleq \sup_{x\in[0,1]} b(x)$ be the best possible return. Since P is unknown, in every trial the learner suffers a loss of $b^* - Y_t$, so that after T trials the learner's regret is

$$R_T \triangleq T\, b^* - \sum_{t=1}^{T} Y_t.$$

With this, return-maximization is the same as regret minimization.

N. Bshouty and C. Gentile (Eds.): COLT 2007, LNAI 4539, pp. 454–468, 2007.

In general, the domain of the decision or action variable X_t can be multi-dimensional. Here we restrict our attention to the one-dimensional case as this shares many of the difficulties of the full multi-dimensional problem, while it allows a simplified presentation of the main ideas.

The continuum-armed bandit problem has many applications (for references see e.g. [1]) and has been studied by a number of authors (e.g., [2,1,3]). It turns out that the continuum-armed bandit problem is much harder than finite-armed bandit problems. For the latter, it is known that logarithmic regret is achievable (see e.g. [4] and the references therein), while for the continuum-armed bandit the regret in typical cases will be polynomial. Concerning results on one-dimensional decision spaces, Kleinberg has derived upper and lower bounds on the regret under the assumption that the mean payoff function is uniformly locally Lipschitz with some exponent $0 < \alpha \leq 1$. Functions in this class satisfy the requirement that there exists some neighborhood size $\delta > 0$ and constant $L \geq 0$ such that for any $x, x' \in [0,1]$ which are δ-close to each other, $|b(x) - b(x')| \leq L|x - x'|^\alpha$ holds. Kleinberg proposed a natural discretization-based algorithm that divides the domain into subintervals of equal lengths and plays a finite-armed bandit problem over the discretized problem. When choosing an interval, Kleinberg's algorithm samples its midpoint. He proves that this algorithm achieves an expected regret of $\tilde{O}\left(T^{2/3}\right)$ over T steps, along with a lower bound of $\Omega(T^{2/3})$ that matches the upper bound apart from a logarithmic factor. If the exponent α is known, the algorithm is shown to achieve expected regret of size $\tilde{O}\left(T^{(1+\alpha)/(1+2\alpha)}\right)$.

In another recent work Cope [3] studies a modified Kiefer-Wolfowitz algorithm (the modification concerns the learning rates). He shows an expected regret bound of size $O(T^{1/2})$ if b is unimodal, three times continuously differentiable, and its derivative is well behaved at its maxima x^* in the sense that $c_1|x - x^*|^2 \leq (x - x^*)b'(x)$ and $|b'(x)| \leq c_2|x - x^*|$ hold for some $c_1, c_2 > 0$.

In this paper, we provide a refined performance characterization for the following modification of Kleinberg's algorithm: While Kleinberg suggested to pick the midpoints of the intervals, we propose to sample actions uniformly at random within the interval. The key underlying idea is the following. There are two sources of the loss in the algorithm: the loss coming from the discretization of the continuous action space (the approximation loss) and the loss for selecting suboptimal arms (cost of learning). A bound on the approximation loss is controlled by the smoothness of the function at its maxima. The cost of learning, on the other hand, is controlled by the gap between the payoffs of suboptimal intervals and the optimal payoff. These gaps are easier to control if one samples uniformly from an interval than if one samples only the midpoint of the interval. Our analysis overcomes another limitation of Kleinberg's analysis which is incapable of capturing higher order smoothness: If b is uniformly locally Lipschitz with coefficient $\alpha > 1$ then it must be constant. We avoid this problem by demanding continuity only at the maxima of the mean payoff function.

A careful analysis then leads to a number of improved bounds. In particular, the modified algorithm achieves expected regret of $\tilde{O}\left(T^{1/2}\right)$ if b has finitely many maxima and non-vanishing, continuous second derivatives at all maxima.

Compared with the result of Cope, the regret is within a logarithmic factor, while our conditions on the payoff function are much weaker. Our upper bounds on the expected regret are complemented by a matching lower bound and a high-probability bound.

2 Problem Setup and Algorithm

In this section we state our assumptions on the mean payoff function, give our algorithm and an outline of the rest of the paper.

Our first assumption is a continuity condition. Without such a condition the regret may grow linearly with T, as it is hard to find maxima of a function, which are obtained at a sharp peak. We propose to capture this difficulty by the degree of continuity at the maxima:

Assumption 1. *There exist constants $L \geq 0$, $\alpha > 0$ such that for any point $x^* \in [0,1]$ with $\limsup_{x \to x^*} b(x) = b^* \triangleq \sup_{x \in [0,1]} b(x)$, and all $x \in [0,1]$*

$$b(x^*) - b(x) \leq L|x^* - x|^\alpha.$$

Define the loss function $d_{x^*}(x) \triangleq b(x^*) - b(x)$. Under Assumption 1, $0 \leq d_{x^*}(x) \leq L|x^* - x|^\alpha$. Hence d_{x^*} is Hölder continuous at x^* with exponent α, and so is b. In particular, $d_{x^*}(x^*) = 0$ and thus $b(x^*) = b^*$. Note that since we do not require this condition to hold at all points in the domain of b, we may allow $\alpha > 1$ without restricting the set of admissible functions to the set of constant functions.

Finding the maximum is also hard, if there are many candidates for the maximum, i.e., if for many x the value of b is close to b^*. This difficulty is captured by the measure of points with value close to the maximum:

Assumption 2. *There exist constants $M \geq 0$, $\beta > 0$ such that for all $\varepsilon > 0$,*

$$m\left(\{\, x \,:\, b^* - \varepsilon < b(x) \leq b^* \,\}\right) \leq M\varepsilon^\beta$$

holds, where m denotes the Lebesgue measure.

In terms of the loss function $d(x) \triangleq b^* - b(x)$ the condition states that $m(\{\, x : d(x) \geq \epsilon \,\}) \geq 1 - M\varepsilon^\beta$. For large β and $\epsilon > 0$, $m(\{\, x : d(x) \geq \epsilon \,\}) \approx 1$. Hence the maxima of the function do not have strong competitors. In fact, Assumptions 1 and 2 are complementary to each other in the sense that $\alpha\beta \leq 1$ holds for most functions. In particular, an elementary argument shows that under these assumptions $\alpha\beta \leq 1$ holds if b is measurable, all maxima of b are in $(0,1)$ and b is not constant in the vicinity of any of its maxima.

Assumptions 1 and 2 put global constraints on the function. We will also consider the following assumption which relaxes this requirement:

Assumption 3. *Let X^* be the set of maxima of b. Then $X^* \subset (0,1)$ and there exist $\rho > 0, \nu > 0, \alpha > 0, \beta > 0$, $L \geq 0, M \geq 0$ such that for any maximum $x^* \in X^*$, Assumptions 1 and 2 hold when x is restricted to the intervals $(x^* - 2\rho, x^* + 2\rho) \subset [0,1]$. Further, it holds that whenever $x \in [0,1] \setminus \bigcup_{x^* \in X^*} (x^* - \rho, x^* + \rho)$ then $b(x) \leq b^* - \nu$.*

Parameter: n
Initialization: Divide $[0,1]$ into n subintervals I_k with $I_k = [\frac{k-1}{n}, \frac{k}{n})$ $(1 \le k < n)$ and $I_n = [\frac{n-1}{n}, 1]$.
Execute UCB on the set of intervals:
- **Initialization:** Choose from each interval I_k a point uniformly at random.
- **Loop:**
  - Choose the interval I_k that maximizes $\hat{b}_k + \sqrt{\frac{2 \ln t}{t_k}}$, where $\hat{b}_k$ is the average return obtained from points in interval I_k, t_k is the number of times interval I_k was chosen, and t is the overall number of steps taken so far.
  - Choose a point uniformly at random from the chosen interval I_k.

Fig. 1. The **UCBC** algorithm with the number of intervals as parameter

This assumption requires that the function is well behaved in the vicinity of its well separated maxima.

As discussed before, we use a discretization-based algorithm that divides the domain into subintervals of equal lengths. Within each subinterval the algorithm chooses the actions uniformly at random. The problem is then to set the number of intervals n and to decide which interval to sample from. While we leave the choice of n open at the moment (n is a parameter of the algorithm, and a central theme of the paper is to find the "right" value of n), for the latter part, just like Kleinberg, we use the UCB algorithm (i.e. UCB1 from [4]). UCB is a finite-armed bandit algorithm that uses upper confidence bounds on the arms' sample-means and achieves optimal logarithmic regret-rates [4]. A more formal description of our UCBC (UCB for continuous bandits) algorithm is given in Figure 1.

Under Assumptions 1 and 2, in Section 3.1 we prove a generic result that gives a bound on the expected regret in terms of the number of subintervals n and the length T of the trial. As will be shown in Section 3.2, this result also holds under Assumption 3. We then give a high probability bound in Section 4. In Section 5, we show that without any knowledge of β, we get the same bounds as Kleinberg. However, for known β we get an improved bound of $\tilde{O}\big(T^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}}\big)$. In particular, if b has finitely many maxima and a non-vanishing, continuous second derivative at all maxima, then we prove $\mathbb{E}[R_T] = \tilde{O}(\sqrt{T})$. We also present lower bounds on the regret under Assumptions 1 and 2 in Section 6. These lower bounds essentially match our upper bound, hence showing that the algorithm's performance is optimal if α, β are known.

3 Bounds on the Expected Regret

3.1 Bounds Under Assumptions 1 and 2

In this section we analyze the regret of UCBC under Assumptions 1 and 2. We use the following result that can be extracted from the analysis of UCB (in particular, from the proof of Theorem 1 in [4]):

Lemma 1. *Consider UCB applied to a multi-armed bandit problem with payoffs in* $[0,1]$*. Let* $\tau_i(T)$ *denote number of times an arm is chosen up to (and including) time step* T*, and let* d_i *be the expected loss when playing arm* i *instead of an optimal arm. If* i *is the index of a suboptimal arm then*

$$\mathbb{E}\left[\tau_i(T)\right] \le \frac{A\ln(T)}{d_i^2} + B \tag{1}$$

for some constants A, B*. In particular, one may select* $A = 8$ *and* $B = 1 + \pi^2/3$*.*

Analysis of the Regret of UCBC: Our analysis will follow the idea described earlier, bounding separately the loss resulting from the discretization, and the cost of learning which interval is the best. According to Lemma 1, for the latter we need to lower bound the gap between the best arm's payoff and the suboptimal arms' payoffs. This is the critical part of the proof.

For $k = 1, 2, \ldots, n$ let I_k denote the k-th interval, i.e. $I_k \triangleq [(k-1)/n, k/n)$ if $1 \le k < n$ and $I_n \triangleq [(n-1)/n, 1]$. Let the choice of UCB be $U_t \in \{1, \ldots, n\}$, the choice of UCBC be X_t and the received payoff Y_t. Let $\tau_i(T) \triangleq \sum_{t=1}^{T} \mathbb{I}_{\{U_t = i\}}$ be the number of times UCBC selects arm i in the first T trials.

Denote by $\overline{b}_k \triangleq n \int_{I_k} b(x)dx$ the expected payoff when the algorithm selects to sample from the k-th subinterval. Let $b_1 \le b_2 \le \ldots \le b_n$ be the ordering of $(\overline{b}_k)_k$, and let π be the permutation that gives this ordering, i.e. $b_{\pi(k)} = \overline{b}_k$. Set $\tau'_i(T) \triangleq \tau_{\pi^{-1}(i)}(T)$. Finally, let $d_i^* \triangleq b^* - b_i$, and $d_i \triangleq b_n - b_i$.

By Wald's identity, the expected regret of UCBC can be expressed via the sampling times $\tau_k(T)$, alternatively using $\tau'_k(T)$, as follows:

$$\mathbb{E}\left[R_T\right] = \sum_{k=1}^{n} (b^* - \overline{b}_k)\mathbb{E}\left[\tau_k(T)\right] = \sum_{i=1}^{n} d_i^* \mathbb{E}\left[\tau'_i(T)\right].$$

In what follows, we analyze $\tilde{R}_T \triangleq \sum_{i=1}^{n} d_i^* \overline{\tau}'_i(T)$, where $\overline{\tau}'_i(T) \triangleq \mathbb{E}\left[\tau'_i(T)\right]$. We start with a simple observation that follows immediately from Assumption 1:

$$d_n^* \triangleq b^* - b_n \le Ln^{-\alpha}. \tag{2}$$

To see that this holds pick any maximum x^* of b and let k^* be the index of the interval containing x^*: $x^* \in I_{k^*}$. Let $i^* \triangleq \pi(k^*)$. Then $b^* - b_n \le b^* - b_{i^*} = b(x^*) - n\int_{I_{i^*}} b(x)dx = n\int_{I_{i^*}} (b(x^*) - b(x))dx \le nL\int_0^{1/n} z^\alpha dz \le nL(1/n)^{\alpha+1}/(\alpha+1) \le Ln^{-\alpha}$ as promised.

We split the set of arms into two parts. Let $\gamma \ge 1$ be a real-valued number to be selected later and define

$$S \triangleq \{\, i \,:\, d_i^* > \gamma Ln^{-\alpha} \,\}.$$

By design, S contains the indices of "strongly" suboptimal intervals.

We split the regret based on if the payoff in an interval is "strongly" suboptimal:

$$\sum_{i=1}^{n} d_i^* \overline{\tau}'_i(T) \le \sum_{i \notin S} d_i^* \overline{\tau}'_i(T) + \sum_{i \in S} d_i^* \overline{\tau}'_i(T) \triangleq \tilde{R}_{T,1} + \tilde{R}_{T,2}.$$

Bounding $\tilde{R}_{T,1}$: $\tilde{R}_{T,1}$ is controlled by the resolution of the discretization: By the choice of S, $d_i^* \le \gamma L n^{-\alpha}$ whenever $i \notin S$. Hence

$$\tilde{R}_{T,1} \le \gamma L n^{-\alpha} \sum_{i \notin S} \overline{\tau}_i'(T) \le \gamma L n^{-\alpha} T. \tag{3}$$

Bounding $\tilde{R}_{T,2}$: The idea here is to "sort" intervals with index in S according to the size of the "gaps" d_i into different buckets and then argue that the number of indices in a bucket with small gaps cannot be too large. Within each bucket we use Lemma 1 to bound the regret.

First, let us note that when $n^\alpha \le \gamma L$ then $S = \emptyset$, hence $\tilde{R}_{T,2} = 0$. Thus, in what follows we assume that $n^\alpha > \gamma L$ or $\gamma L n^{-\alpha} < 1$.

Observe that S does not contain any interval with an optimal response: if $b_i = b_n$, then $i \notin S$. Indeed, by (2) $d_n^* \le L n^{-\alpha} \le \gamma L n^{-\alpha}$. Therefore, we may use Lemma 1 to bound $\overline{\tau}_i'(T)$ for $i \in S$. By (2), $d_i^* = b^* - b_i \le b_n - b_i + L n^{-\alpha} = d_i + L n^{-\alpha}$ and hence using (1) we get

$$\tilde{R}_{T,2} \le A \ln(T) \sum_{i \in S} \left(\frac{1}{d_i} + \frac{L n^{-\alpha}}{d_i^2} \right) + B|S|. \tag{4}$$

Let $\Delta_k \triangleq 2^{-k}$, $k = 0, 1, 2, \ldots$ so that $1 = \Delta_0 > \Delta_1 > \Delta_2 > \ldots$. Let

$$S_k \triangleq \{\, i \in S \,:\, \Delta_k \le d_i^* < \Delta_{k-1} \,\}, \quad k = 0, 1, 2, \ldots .$$

Note that if $\Delta_{k-1} \le \gamma L n^{-\alpha}$ then $S_k = \emptyset$. Hence, if we define K to be the unique index such that $\gamma L n^{-\alpha} \in [\Delta_K, \Delta_{K-1})$, then $S = \bigcup_{k=0}^K S_k$. (The existence of K is guaranteed since by assumption $\gamma L n^{-\alpha} < 1$.) Note that $K = \lceil \ln_2(n^\alpha/(\gamma L)) \rceil$, and if $k \le K$, then $\Delta_{k-1} > \gamma L n^{-\alpha}$. Now set $\gamma \triangleq 4$. By (2), $d_i \triangleq b_n - b_i \ge b^* - L n^{-\alpha} - b_i = d_i^* - L n^{-\alpha}$, hence for $i \in S_k$, $k = 0, 1, \ldots, K$,

$$d_i \ge \Delta_k - L n^{-\alpha} = \Delta_k (1 - L n^{-\alpha}/\Delta_k) > \Delta_k/2. \tag{5}$$

Here in the last step we used that $\Delta_k = (1/2)\,\Delta_{k-1} > (1/2)\,\gamma L n^{-\alpha} = 2 L n^{-\alpha}$. Using (5) we get

$$\sum_{i \in S} \left(\frac{1}{d_i} + \frac{L n^{-\alpha}}{d_i^2} \right) = \sum_{k=0}^K \sum_{i \in S_k} \left(\frac{1}{d_i} + \frac{L n^{-\alpha}}{d_i^2} \right) \le \sum_{k=0}^K \left(\frac{2}{\Delta_k} + \frac{4 L n^{-\alpha}}{\Delta_k^2} \right) |S_k|. \tag{6}$$

A Bound on $|S_k|$: Let $U_i(\varepsilon) \triangleq \{\, x \in I_i \,:\, b^* - b(x) \ge \varepsilon \,\}$ with some $\varepsilon > 0$. Note that $b^* - b(x) \ge 0$ and hence by Markov's inequality, $m(U_i(\varepsilon)) \le (1/\varepsilon) \int_{I_i} (b^* - b(x)) dx = (b^* - b_i) m(I_i)/\varepsilon = d_i^* m(I_i)/\varepsilon$ and thus for $\overline{U}_i(\varepsilon) = I_i \setminus U_i(\varepsilon)$, $m(\overline{U}_i(\varepsilon)) \ge (1 - d_i^*/\varepsilon)\, m(I_i)$. Assume that $i \in S_k$. By the definition of S_k, $\Delta_{k-1} > d_i^*$ and hence $m(\overline{U}_i(\varepsilon)) \ge (1 - \Delta_{k-1}/\varepsilon)\, m(I_i)$. Set $\varepsilon = 2\Delta_{k-1}$ so that $m(\overline{U}_i(2\Delta_{k-1})) \ge 1/2\, m(I_i)$. Therefore,

$$|S_k| m(I_1) = \sum_{i \in S_k} m(I_i) \le 2 \sum_{i \in S_k} m(\overline{U}_i(2\Delta_{k-1})) = 2\, m(\cup_{i \in S_k}^* \overline{U}_i(2\Delta_{k-1})), \tag{7}$$

where the disjointness follows since $\overline{U}_i(2\Delta_{k-1}) \subset I_i$. Since $\overline{U}_i(2\Delta_{k-1}) = \{\, x \in I_i \,:\, b^* - b(x) \le 2\Delta_{k-1} \,\} = \{\, x \in I_i \,:\, b(x) \ge b^* - \Delta_{k-2} \,\}$, the union of these sets is contained in $\{\, x \in [0,1] \,:\, b(x) \ge b^* - \Delta_{k-2} \,\}$ and therefore by Assumption 2, $m(\bigcup_{i\in S_k}^{*} \overline{U}_i(2\Delta_{k-1})) \le M\,(4\,\Delta_k)^\beta$. Combined with (7), this gives $|S_k| m(I_1) \le 2M\,(4\,\Delta_k)^\beta$ and hence $|S_k| \le 2\,Mn\,(4\,\Delta_k)^\beta$.

Putting Things Together: The bound on $|S_k|$ together with (6) and (4) yields

$$\begin{aligned}\tilde{R}_{T,2} &\le 2AMn\ln(T)\left(\sum_{k=0}^{K}(4\,\Delta_k)^\beta\left(\frac{2}{\Delta_k}+\frac{4Ln^{-\alpha}}{\Delta_k^2}\right)\right)+B\,n\\ &= 4^{\beta+1}AMn\ln(T)\left(\sum_{k=0}^{K}2^{(1-\beta)k}+2Ln^{-\alpha}\sum_{k=0}^{K}2^{(2-\beta)k}\right)+B\,n. \qquad (8)\end{aligned}$$

Assuming that $\beta \notin \{0,1,2\}$ and exploiting that $2^{K+1} \le n^\alpha/L$ (this follows from $K-1 \le \ln_2(1/(\gamma L n^{-\alpha}))$ and $\gamma = 4$), we get

$$\tilde{R}_{T,2} \le 4^{\beta+1}AMn\ln(T)\left(\frac{(n^\alpha/L)^{1-\beta}-1}{2^{1-\beta}-1}+2Ln^{-\alpha}\frac{(n^\alpha/L)^{2-\beta}-1}{2^{2-\beta}-1}\right)+Bn. \quad (9)$$

Considering β:

- If $\beta < 1$, from (9) we get via some tedious calculations,

$$\tilde{R}_{T,2} \le \frac{3\cdot 4^{\beta+1}\,AML^{\beta-1}}{2^{1-\beta}-1}\,n^{1+\alpha-\alpha\beta}\ln(T)+Bn.$$

- $\beta = 1$: Since by our earlier remark we assume that $n^\alpha > \gamma L = 4L$, working directly from (8) gives

$$\tilde{R}_{T,2} \;\le\; \frac{4^{\beta+1}AM\alpha}{\ln 2}\,n\,\ln n\,\ln T + 4^{\beta+1}AM(3+\ln_2(2/L))\,n\ln(T)+Bn.$$

- $1 < \beta < 2$: Using (9) and $n^\alpha > \gamma L > L$ we get

$$\tilde{R}_{T,2} \le \frac{4^{\beta+1}AM}{1-2^{1-\beta}}n\ln T+\frac{2\,4^{\beta+1}AML^{\beta-1}}{2^{2-\beta}-1}n^{1+\alpha-\alpha\beta}\ln T+Bn.$$

- If $\beta = 2$, from (8) using $\ln x/x \le 1/e$ we get

$$\tilde{R}_{T,2} \le 2\,4^{\beta+1}\left(1+\frac{1}{4e\ln 2}\right)AMn\ln T+2\,4^{\beta+1}AMLn^{1-\alpha}\ln T+Bn.$$

- If $\beta > 2$, using again (9),

$$\tilde{R}_{T,2} \le \frac{4^{\beta+1}AM}{1-2^{1-\beta}}\,n\ln T+\frac{2\,4^{\beta+1}AML}{1-2^{2-\beta}}\,n^{1-\alpha}\ln T+Bn.$$

Combining these inequalities with the bound (3) on $\tilde{R}_{T,1}$, we get:

Lemma 2. *Consider UCBC with n intervals in a continuum-armed bandit problem where the payoffs are in the range $[0,1]$ and the mean payoff function satisfies Assumptions 1 and 2 with some constants L, α, M, β. Then*

$$\mathbb{E}[R_T] \le 4Ln^{-\alpha}T + 4^{\beta+1}AMnR'_T \ln T + B\,n,$$

where

$$R'_T = \begin{cases} \frac{3\,L^{\beta-1}}{2^{1-\beta}-1}\, n^{\alpha-\alpha\beta}, & 0 \le \beta < 1; \\ \frac{\alpha}{\ln 2}\ \ln n + 3 + \ln_2(2/L), & \beta = 1; \\ \frac{1}{1-2^{1-\beta}} + \frac{2L^{\beta-1}}{2^{2-\beta}-1} n^{\alpha-\alpha\beta}, & 1 < \beta < 2; \\ 2\left(1 + \frac{2}{4e\ln 2}\right) + 2L, & \beta = 2; \\ 2 + \frac{2L}{1-2^{2-\beta}}, & \beta > 2. \end{cases}$$

Here A, B are as in Lemma 1 and can be selected as $A = 8$, $B = 1 + \pi^2/3$.

3.2 Bounds Under the Localized Assumption (Assumption 3)

The previous analysis can be repeated, except that we split S into two disjoint parts: $S' \triangleq \{\, i \in S \,:\, I_i \cap (xs-\rho, xs+\rho) \neq \emptyset \,\}$, $S'' \triangleq \{\, i \in S \,:\, I_i \cap (xs-\rho, xs+\rho) = \emptyset \,\}$. If n is big enough so that $1/n < \rho$, we can use the argument of the previous section for S', since then Assumptions 1 and 2 hold for any interval in S'. For $i \in S''$, by Assumption 3, $d_i \ge \nu$. Hence, $d_i^* d_i^{-2} \le 1/\nu + Ln^{-\alpha}/\nu^2$, so that

$$\sum_{i \in S''} d_i^* d_i^{-2} \le \frac{n}{\nu} + \frac{Ln^{1-\alpha}}{\nu^2} \le \frac{n}{\nu}\left(1 + \frac{L}{\nu}\right).$$

Let $c \triangleq \frac{1}{\nu}\left(1 + \frac{L}{\nu}\right)$. Then we get the following result:

Lemma 3. *Under Assumption 3, the expected regret satisfies*

$$\mathbb{E}[R_T] \le 4Ln^{-\alpha}T + 4^{\beta+1}AMnR'_T \ln T + (B\ + c)n.$$

4 A High Probability Bound

The next lemma follows from a version of Bernstein's inequality due to Cesa-Bianchi et al. [5]:

Lemma 4. *Let $J_s \in \mathcal{F}_{s-1}$, $J_s \in \{0,1\}$, $Z_s \in \mathcal{F}_s$, $|Z_s| \le K$, and assume that $\mathbb{E}[J_s Z_s | \mathcal{F}_{s-1}] = 0$. Let $M_t \triangleq \sum_{s=1}^{t} J_s Z_s$, $T_t \triangleq \sum_{s=1}^{t} J_s$. Then for all $\delta > 0$, $t > 0$,*

$$\mathbb{P}\left(M_t > K\phi(t, T_t, \delta)\right) \le \delta,$$

where

$$\phi(t, T, \delta) \triangleq \sqrt{2(T_t + 1)\ln(t/\delta)} + \tfrac{\sqrt{2}}{3}\ \ln(t/\delta).$$

Proof. Let $X_s = J_s Z_s$. Then X_s is a bounded martingale difference series. Observe that $\mathbb{E}\left[X_s^2 \,|\, \mathcal{F}_{s-1}\right] \le K^2 J_s$ and hence $V_t = \sum_{s=1}^t \mathbb{E}\left[X_s^2 \,|\, \mathcal{F}_{s-1}\right] \le K^2 T_t$. Hence,

$$\mathbb{P}\left(M_t \ge \sqrt{2(K^2 T_t + K^2)\ln(t/\delta)} + \tfrac{\sqrt{2}}{3} K \ln(t/\delta)\right) \le$$
$$\mathbb{P}\left(M_t \ge \sqrt{2(V_t + K^2)\ln(t/\delta)} + \tfrac{\sqrt{2}}{3} K \ln(t/\delta)\right) \le \delta,$$

where the last inequality follows by Corollary 16 of [5]. □

Now consider UCBC on a bandit problem defined by P, but let us now change the protocol of interaction. In particular, let $(X_{k,t}, Y_{k,t})_{k,t}$ be a sequence of random variables generated independently of the choice of the algorithm such that $X_{k,t}$ is uniformly distributed in the interval I_k and $Y_{k,t} \sim P(\cdot|X_{k,t})$. We assume that these samples are generated independently of each other and from the past. Remember that Y_t is the payoff received at time step t, and let U_t again denote the index of the interval chosen at time step t, $Y'_t = Y_{U_t,t}$, $X'_t = X_{U_t,t}$. It should be clear, that the distributions of $\sum_{t=1}^T Y_t$ and $\sum_{t=1}^T Y'_t$ are identical. Hence, it suffices to analyze the properties of the regret of UCBC when it is used under this new protocol. For simplicity, in what follows we will use Y_t instead of Y'_t and X_t instead of X'_t.

Fix any index $i \in \{1, 2, \ldots, n\}$. Remember that $d_i^* \triangleq b^* - b_i$ is the expected loss when playing interval i, where b_i is the expected payoff for interval i and $b_i \le b_{i+1}$, $i = 1, \ldots, n-1$. Let $\mathcal{F}_s = \sigma(X_1, X_{i,1}, Y_1, Y_{i,1}, \ldots, X_s, X_{i,s}, Y_s, Y_{i,s})$, $Z_s = b^* - Y_{i,s} - d_i^*$, $J_s = \mathbb{I}_{\{U_s = i\}}$. It can be readily verified that the conditions of Lemma 4 are satisfied with $K = 1$. Hence, with probability at least $1 - \delta/(2n)$ simultaneously for all i, $\phi(T, \tau_i(T), \delta/(2n)) \ge \sum_{s=1}^T J_s Z_s = \sum_{s=1}^T \mathbb{I}_{\{U_t = i\}}(b^* - Y_{i,s} - d_i^*)$, i.e.,

$$\sum_{s=1}^T \mathbb{I}_{\{U_t=i\}}(b^* - Y_{i,s}) \le d_i^* \sum_{s=1}^T \mathbb{I}_{\{U_t=i\}} + \phi(T, \tau_i(T), \delta/(2n)). \tag{10}$$

Summing (10) over i, followed by some calculations gives $R_T \le \sum_{i=1}^n d_i^* \tau_i(T) + H_T(\delta)$, where

$$H_T(\delta) \triangleq \left(\sqrt{Tn} + n\right)\sqrt{2\ln(2Tn/\delta)} + 2(\sqrt{2}/3) n \ln(2Tn/\delta).$$

Our aim now is to obtain a high probability upper bound on $\tau_i(T)$ for the suboptimal arms. For this we change the confidence intervals of the algorithm to $c_{t,s}(\delta_0) = \sqrt{\frac{2\ln(t/\delta_0)}{s}}$, i.e. the modified UCBC algorithm (called UCBC(δ_0)) chooses the interval that maximizes $\hat{b}_k + c_{t,t_k}(\delta_0)$ for an appropriately selected δ_0. Consider inequality (6) in the proof of Theorem 1 in [4]. Using the notation of [4], the expectation over the triple sum in inequality (6) is $O(\delta_0)$, and thus the probability that $T_i > \ell_i + \Omega(1)$, $\ell_i \triangleq 8\ln(n/\delta_0)/\Delta_i^2$, is $O(\delta_0)$ by Markov's inequality. Hence, the following result holds:

Lemma 5. *Under the assumptions of Lemma 1, with probability at least* $1-n\delta_0$*, simultaneously for all suboptimal arms* i*,*

$$\tau_i(T) \le \frac{A' \ln(T/\delta_0)}{d_i^2} + B'$$

for some constants A', B'*.*

Setting $\delta_0 \triangleq \delta/(2n)$ in UCBC(δ_0), we get that $\tau_i(T) \le A' \ln(2Tn/\delta)/d_i^2 + B'$ holds for all suboptimal arms simultaneously with probability at least $1-\delta/2$. Hence, with probability at least $1-\delta$,

$$\begin{aligned} R_T &\le 4Ln^{-\alpha}T + \sum_{i \in S} d_i^* \tau_i(T) + H_T(\delta/2) \\ &\le 4Ln^{-\alpha}T + A' \ln(2Tn/\delta)\Big(\sum_{i \in S} d_i^* d_i^{-2}\Big) + nB' + H_T(\delta/2). \end{aligned}$$

Continuing as in Section 3, we obtain the following result:

Lemma 6. *Let* $\delta > 0$*. Consider UCBC with* n *intervals and confidence sequence* $c_{t,s}(\delta/(2n))$ *applied to a continuum-armed bandit problem, where the payoffs are in the range* $[0,1]$ *and the mean payoff function satisfies Assumptions 1 and 2 with some constants* L, α, M, β*. If* $n^\alpha \le 4L$*, then the regret satisfies* $R_T \le 4Ln^{-\alpha}T + H_T(\delta/2)$ *with probability* $1-\delta$*, while for* $n^\alpha > 4L$ *it holds with probability at least* $1-\delta$ *that*

$$R_T \le 4Ln^{-\alpha}\, T + 4^{\beta+1} A' M n R'_T \ln(2Tn/\delta) + B'\, n + H_T(\delta/2).$$

Using the reasoning of Section 3.2, this result can be extended to the localized version of Assumptions 1 and 2 (Assumption 3). We omit the details.

5 Choice of the Parameters

First note that according to Lemma 2 we have for $0 \le \beta < 1$ and a suitable constant c

$$\mathbb{E}\,[R_T] \le 4L\frac{T}{n^\alpha} + \frac{c\, 4^\beta\, M L^{\beta-1}}{2^{1-\beta}-1} n^{1+\alpha-\alpha\beta} \ln T. \tag{11}$$

5.1 Results Without Assumption 2

With $\beta = 0$ and $M = 1$ Assumption 2 trivially holds true. From (11) we get

$$\mathbb{E}\,[R_T] \le 4L\frac{T}{n^\alpha} + \frac{c}{L} n^{1+\alpha} \ln T.$$

Corollary 1. *If α is known, setting $n \triangleq \left(\frac{T}{\ln T}\right)^{\frac{1}{1+2\alpha}}$ gives*

$$\mathbb{E}\left[R_T\right] \leq \left(4L + \frac{c}{L}\right) T^{\frac{1+\alpha}{1+2\alpha}} (\ln T)^{\frac{\alpha}{1+2\alpha}}, \tag{12}$$

while if α is unknown, setting $n \triangleq \left(\frac{T}{\ln T}\right)^{\frac{1}{3}}$ gives for sufficiently large T

$$\mathbb{E}\left[R_T\right] \leq 4L \cdot T^{\max\{1-\frac{\alpha}{3}, \frac{2}{3}\}} (\ln T)^{\frac{1}{3}} + \frac{c}{L} \cdot T^{\frac{2}{3}} (\ln T)^{\frac{2}{3}}. \tag{13}$$

Proof. (12) is straightforward. Concerning (13), first note that for our choice of n and $\alpha \leq 1$ we have

$$\mathbb{E}\left[R_T\right] \;\leq\; 4L \cdot T^{1-\frac{\alpha}{3}} (\ln T)^{\frac{1}{3}} + \frac{c}{L} \cdot T^{\frac{2}{3}} (\ln T)^{\frac{2-\alpha}{3}}. \tag{14}$$

On the other hand, if $\alpha > 1$, then $Ln^{-\alpha} \leq L\sqrt{nT^{-1} \ln T}$ for $n = \left(\frac{T}{\ln T}\right)^{1/3}$. Then $\mathbb{E}\left[R_T\right] \;\leq\; \left(4L + \frac{c}{L}\right) T^{\frac{2}{3}} (\ln T)^{\frac{1}{3}}$. Combining this with (14) gives (13). □

5.2 Results Using Assumption 2

The most interesting case is $\beta < 1$. For known α and β, we set $n \triangleq \left(\frac{T}{\ln T}\right)^{\frac{1}{1+2\alpha-\alpha\beta}}$ and get from (11), $\mathbb{E}\left[R_T\right] \leq \left(4L + \frac{4cML^{\beta-1}}{2^{1-\beta}-1}\right) \cdot T^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}} (\ln T)^{\frac{\alpha}{1+2\alpha-\alpha\beta}}$.

As noted before, comparing Assumptions 1 and 2 we find that for most functions b we have $\alpha\beta \leq 1$, the only exception being when b is constant in the vicinity of the maximum. Making the optimistic assumption that $\alpha\beta = 1$, we may set $n \triangleq \left(\frac{T}{\ln T}\right)^{\frac{1}{2\alpha}}$ and get

$$\mathbb{E}\left[R_T\right] \leq \left(4L + \frac{4cML^{\beta-1}}{2^{1-\beta}-1}\right) \cdot \sqrt{T \ln T}. \tag{15}$$

If the function b has continuous second derivatives, then Assumptions 1 and 2 are satisfied with $\alpha = 2$ and $\beta = 1/2$:

Theorem 1. *If b has a finite number of maxima x^* with $\limsup_{x \to x^*} b(x) = b^*$, and continuous second derivatives $\neq 0$ at all these x^*, then our algorithm with $n \triangleq \left(\frac{T}{\ln T}\right)^{\frac{1}{4}}$ achieves*

$$\mathbb{E}\left[R_T\right] \leq O\left(\sqrt{T \ln T}\right).$$

Proof. By assumption, $b'(x^*) = 0$ and $b''(x^*) \neq 0$ for any maximum x^*. Using Taylor series expansion we find

$$b(x^*) - L_1(x^* - x)^2 \leq b(x) \leq b(x^*) - L_2(x^* - x)^2 + L_3|x^* - x|^3$$

for suitable constants $L_1, L_2, L_3 > 0$, any maximum x^*, and any $x \in [0, 1]$. Hence, Assumption 1 is satisfied with $\alpha = 2$.

Furthermore, there are $\varepsilon_0 > 0$ and $0 < \delta_0 < L_2/(2L_3)$ such that $b(x) \le b^* - \varepsilon_0$ for all x with $\min_{x^*} |x - x^*| \ge \delta_0$. Thus $b(x) > b(x^*) - \varepsilon$ for $\varepsilon < \varepsilon_0$ implies $\min_{x^*} |x - x^*| < \delta_0$ and $b(x^*) - \varepsilon < b(x) \le b(x^*) - L_2(x^* - x)^2 + L_3|x^* - x|^3 = b(x^*) - (x^* - x)^2(L_2 - L_3|x^* - x|) \le b(x^*) - L_2(x^* - x)^2/2$ such that $|x - x^*| < \sqrt{2\varepsilon/L_2}$ for some maximum x^* (out of the finitely many). For $\varepsilon \ge \varepsilon_0$ we have $|x - x^*| \le 1 \le \sqrt{\varepsilon/\varepsilon_0}$. Hence, Assumption 2 is satisfied with $\beta = 1/2$. The theorem follows from (15). □

5.3 When the Number of Steps T Is Unknown

If unlike in the previous sections the total number of steps T is unknown, then a simple application of the doubling trick gives the same bounds with somewhat worse constants. That is, UCBC is executed for 2^k steps in rounds $k = 1, 2, \ldots$. Then after T steps at most $K = 1 + \lceil \ln_2 T \rceil$ rounds have been played. Thus the total regret can be obtained by summing up over all rounds $\sum_{k=1}^{K} (2^k)^a (\ln(2^k))^b = O\left((2^K)^a (\ln(2^K))^b\right) = O\left(T^a (\ln T)^b\right)$.

6 Lower Bounds on the Regret

In this section we extend the lower bound result of Kleinberg [1] and show that our upper bounds on the regret are tight (apart from a logarithmic factor).

Theorem 2. *For any $\alpha > 0, \beta \ge 0$, $\alpha\beta \le 1$, and any learning algorithm, there is a function b satisfying Assumptions 1 and 2 such that for any $\gamma < \frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}$,*

$$\limsup_{T \to \infty} \frac{\mathbb{E}[R_T]}{T^\gamma} \to \infty.$$

In [1] this theorem was proven for $\beta = 0$. We extend the construction of [1] to consider also $\beta > 0$.

Proof. We define function b as

$$b(x) \triangleq \sum_{k=k_0}^{\infty} [\phi_k(x) + \psi_k(x)]$$

for an appropriate k_0 and functions ϕ_k and ψ_k. We set $c_{k_0-1} = 0$ and $d_{k_0-1} = 1$ and iteratively define intervals $[c_k, d_k]$ at random. The functions ϕ_k and ψ_k are defined in respect to these random intervals. As such, the function b is constructed by a random process. We will argue, that for any learning algorithm the average regret in respect to this random process is large, which will imply the theorem.

The functions ϕ_k and ψ_k are continuous, non-negative, and positive only within a part of the interval $[c_{k-1}, d_{k-1}]$. The main part of these functions is a plateau where they remain constant, and they rise to and fall from this plateau governed by a function $f : [0,1] \mapsto [0,1]$ where $f(x) \triangleq 1 - (1-x)^\alpha$, such

that $f(0) = 0$ and $f(1) = 1$. The lengths δ_k of the intervals $[c_k, d_k]$ are very rapidly decreasing. We are also using sub-intervals $[c'_k, d'_k]$ with the property $[c_k, d_k] \subset [c'_{k-1}, d'_{k-1}] \subset [c_{k-1}, d_{k-1}]$. Let

$$\delta_k \triangleq 2^{-k!}, \qquad \Delta_k \triangleq \frac{1}{5}\delta_{k-1}^{1/(\alpha\beta)}, \qquad L_k \triangleq \max\left\{2, \left\lfloor \tfrac{1}{5}\delta_k^{\alpha\beta-1}\delta_{k-1} \right\rfloor\right\},$$
$$c'_{k-1} \triangleq c_{k-1} + \Delta_k, \quad \text{and} \quad d'_{k-1} \triangleq c_{k-1} + \Delta_k + 3L_k\delta_k,$$

and

$$\phi_k(x) \triangleq \begin{cases} 0 & \text{for} \quad x \le c_{k-1} \\ \Delta_k^\alpha f(\frac{x-c_{k-1}}{\Delta_k}) & \text{for} \quad c_{k-1} \le x \le c'_{k-1} \\ \Delta_k^\alpha & \text{for} \quad c'_{k-1} \le x \le d'_{k-1} \\ \Delta_k^\alpha f(1 - \frac{x-d'_{k-1}}{\Delta_k}) & \text{for} \quad d'_{k-1} \le x \le d'_{k-1} + \Delta_k \\ 0 & \text{for} \quad d'_{k-1} + \Delta_k \le x \end{cases}.$$

Observe that

$$d'_{k-1} + \Delta_k \le c_{k-1} + 2\Delta_k + 3L_k\delta_k \;\le\; c_{k-1} + \frac{2}{5}\delta_{k-1}^{1/(\alpha\beta)} + \max\left\{6\delta_k, \frac{3}{5}\delta_k^{\alpha\beta}\delta_{k-1}\right\}$$
$$\le c_{k-1} + \frac{2}{5}\delta_{k-1} + \frac{3}{5}\delta_{k-1} \;\le\; c_{k-1} + \delta_{k-1}\,.$$

Let $\ell_k \in \{0, \ldots, L_k - 1\}$ be chosen uniformly at random and set

$$c_k \triangleq c'_{k-1} + (L_k + \ell_k)\delta_k, \quad d_k \triangleq c_k + \delta_k,$$

and

$$\psi_k(x) \triangleq \begin{cases} 0 & \text{for} \quad x \le c_k - \delta_k \\ \delta_k^\alpha f(\frac{x-c_k+\delta_k}{\delta_k}) & \text{for} \quad c_k - \delta_k \le x \le c_k \\ \delta_k^\alpha & \text{for} \quad c_k \le x \le d_k \\ \delta_k^\alpha f(1 - \frac{x-d_k}{\delta_k}) & \text{for} \quad d_k \le x \le d_k + \delta_k \\ 0 & \text{for} \quad d_k + \delta_k \le x \end{cases}.$$

Then any fixed b has a unique maximum at $x^* = \lim_k c_k = \lim_k d_k$. The intuition of the construction is the following: the slope of function f is responsible for matching Assumption 1 tightly (this is rather obvious), whereas the length $3L_k\delta_k$ of $[c'_{k-1}, d'_{k-1}]$ is responsible for matching Assumption 2 tightly. This can be seen from the fact that the peak of function b on top of the plateau $[c'_k, d'_k]$ is approximately of size $\varepsilon = \delta_k^\alpha$, such that $L_k\delta_k \approx \delta_k^{\alpha\beta}\delta_{k-1} \approx \varepsilon^\beta$. ($\delta_{k-1}$ is very large compared to δ_k and can be ignored.).

The heights of functions ϕ_k and ψ_k are chosen such that Assumptions 1 and 2 are satisfied. We first check that function b satisfies Assumption 1. For any $x \in [0, 1]$, $x \neq x^*$, there is a $k \ge 1$ such that $x \in [c_{k-1}, d_{k-1}] \setminus [c_k, d_k]$. We assume without loss of generality that $x < c_k < x^*$. Then $b(x^*) - b(x) = b(x^*) - b(c_k) + b(c_k) - b(x)$ and

$$b(x^*) - b(c_k) \le \sum_{i=k+1}^{\infty} (\Delta_i^\alpha + \delta_i^\alpha) \le 2(\Delta_{k+1}^\alpha + \delta_{k+1}^\alpha) \le 4(x^* - c_k)^\alpha,$$

since $x^* - c_k > \Delta_{k+1} + \delta_{k+1}$. To bound $b(c_k) - b(x)$ consider the following cases:

a) If $c_k - \delta_k \le x < c_k$, then $b(c_k) - b(x) = \delta_k^\alpha - \delta_k^\alpha f\left(\frac{x-c_k+\delta_k}{\delta_k}\right) = \delta_k^\alpha - \delta_k^\alpha\left(1 - \left(1 - \frac{x-c_k+\delta_k}{\delta_k}\right)^\alpha\right) = (c_k - x)^\alpha$.
b) If $c'_{k-1} \le x \le c_k - \delta_k$, then $b(c_k) - b(x) \le \delta_k^\alpha \le (c_k - x)^\alpha$.
c) If $c_{k-1} \le x \le c'_{k-1}$, then $b(c_k) - b(x) = \delta_k^\alpha + \Delta_k^\alpha[1 - f(\frac{x-c_{k-1}}{\Delta_k}] = \delta_k^\alpha + \Delta_k^\alpha[1 - \frac{x-c_{k-1}}{\Delta_k}]^\alpha = \delta_k^\alpha + \Delta_k^\alpha[\frac{c'_{k-1}-x}{\Delta_k}]^\alpha \le 2(c_k - x)^\alpha$.

Since $4(x^* - c_k)^\alpha + 2(c_k - x)^\alpha \le 6(x^* - x)^\alpha$, Assumption 1 is satisfied.

For checking Assumption 2, let $x \in [0, 1]$ be such that $b(x^*) - b(x) < \varepsilon$. We distinguish two cases, $\delta_k^\alpha \le \varepsilon < \Delta_k^\alpha$ and $\Delta_k^\alpha \le \varepsilon < \delta_{k-1}^\alpha$ for some $k \ge k_0$.

a) If $\delta_k^\alpha \le \varepsilon < \Delta_k^\alpha$, then $\phi_k(x) \ge \Delta_k^\alpha + \delta_k^\alpha - \varepsilon$, which by definition of ϕ_k and f holds if $x \in [c'_{k-1} - (\varepsilon - \delta_k^\alpha)^{1/\alpha}, d'_{k-1} + (\varepsilon - \delta_k^\alpha)^{1/\alpha}]$. As $(\varepsilon - \delta_k^\alpha)^{1/\alpha} < \varepsilon^{1/\alpha} \le \varepsilon^\beta$ and $d'_{k-1} - c'_{k-1} = 3L_k\delta_k \le \frac{3}{5}\delta_k^{\alpha\beta}\delta_{k-1} \le \delta_k^{\alpha\beta} \le \varepsilon^\beta$, the length of this interval does not exceed $3\varepsilon^\beta$.
b) On the other hand, if $\Delta_k^\alpha \le \varepsilon < \delta_{k-1}^\alpha$, then $\psi_{k-1}(x) \ge \Delta_k^\alpha + \delta_{k-1}^\alpha - \varepsilon$, which by definition of ψ_{k-1} and f holds, if $x \in [c_{k-1} - (\varepsilon - \Delta_k^\alpha)^{1/\alpha}, d_{k-1} + (\varepsilon - \Delta_k^\alpha)^{1/\alpha}]$. The length of this interval is smaller than $7\varepsilon^\beta$, since $(\varepsilon - \Delta_k^\alpha)^{1/\alpha} < \varepsilon^\beta$ and $d_{k-1} - c_{k-1} = \delta_{k-1} = 5\Delta_k^{\alpha\beta} \le 5\varepsilon^\beta$.

Thus Assumption 2 is satisfied, too.

Finally we show that for $T_k \triangleq \left\lfloor \frac{1}{2}L_k\delta_k^{-2\alpha} \right\rfloor = \Theta\left(\delta_k^{\alpha\beta-1-2\alpha}\delta_{k-1}\right)$, and any $\gamma < \frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}$

$$\lim_{k\to\infty} \frac{\mathbb{E}\left[R_{T_k}\right]}{T_k^\gamma} \to \infty \ .$$

For any $x_1, x_2 \in [c'_{k-1}, d'_{k-1}]$, the Kullback-Leibler distance between the bandits x_1 and x_2 is $O(\delta_k^{2\alpha})$ and it is 0 if both $x_1, x_2 \notin [c_k, d_k]$. Therefore, to identify $[c_k, d_k]$ with probability $\Omega(1)$, at least half of the intervals $[c'_{k-1} + (L_k + \ell_k)\delta_k, c'_{k-1} + (L_k + \ell_k + 1)\delta_k]$, $\ell_k \in 0, \ldots, L_k - 1$, need to be probed $\lceil \delta_k^{-2\alpha} \rceil$ times. Since $b(x^*) - b(x) \ge \delta_k^\alpha$ for $x \notin [c_k, d_k]$, we find

$$\begin{aligned}\mathbb{E}\left[R_{T_k}\right] &= \Omega\left(T_k\delta_k^\alpha\right) \ = \ \Omega\left(\delta_k^{\alpha\beta-1-\alpha}\delta_{k-1}\right) \\ &= \Omega\left(T_k^{\frac{\alpha\beta-1-\alpha}{\alpha\beta-1-2\alpha}}\delta_{k-1}^{1-\frac{\alpha\beta-1-\alpha}{\alpha\beta-1-2\alpha}}\right) \ = \ \Omega\left(T_k^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}}\delta_{k-1}^{\frac{\alpha}{1+2\alpha-\alpha\beta}}\right) \ .\end{aligned}$$

Since $\lim_{k\to\infty} \delta_{k-1}^{\gamma_1}/\delta_k^{\gamma_2} \to \infty$ for any $\gamma_1, \gamma_2 > 0$, this proves the theorem. □

7 Conclusions and Future Work

We have shown that by changing the algorithm of Kleinberg, it is possible to get improved regret bounds under a wide range of conditions. In particular, the uniform local Lipschitz condition is replaced with a smoothness condition that is localized to the set of maxima of the payoff function. A complementary

condition ensures that the maxima do not have many strong competitors. These two conditions allow us to get improved bounds compared to the bounds of Kleinberg [1]. Moreover, the new algorithm is shown to match the performance of the Kiefer-Wolfowitz algorithm [3], but under substantially weaker conditions.

One limitation of the presented results is that in order to get the best possible rates, the user must know the exponents α, β of Assumptions 1 and 2. It is an open question, if it is possible to achieve the optimal rates when this knowledge is not available, possibly by restricting the payoff function in some other reasonable way. In connection to this, we could recently show that a two-phase algorithm achieves regret of $\tilde{O}\left(T^{1/2}\right)$ for functions with well separated maxima, provided that in a small neighborhood of the maxima the functions are unimodal and satisfy a not too strong rate-condition (which holds e.g. for locally strictly convex functions). There is also room for improvement regarding the high probability bounds. Thus, a better bound on the inferior sampling time for UCB would immediately lead to better bounds for our algorithm.

We have not considered d-dimensional action spaces in this paper, though we believe that our results can be extended to this case. Previous lower bounds show that in the worst-case, the regret would scale exponentially with the dimension d. An interesting open question is if there exists an algorithm that scales better when the mean payoff function depends only on some unknown subset of the variables.

Finally, let us remark that if UCB is replaced with UCB-tuned [4] and there is no observation noise, i.e. $Y_t = b(X_t)$, then the analysis presented can be used to prove the improved rate $\tilde{O}\left(T^{1/(1+\alpha)}\right)$, i.e. $\tilde{O}\left(T^{1/2}\right)$ for $\alpha = 1$. Hence, the cost of control-learning is substantially less when there is no observation noise present.

Acknowledgements. Csaba Szepesvári greatly acknowledges the support received through the Alberta Ingenuity Center for Machine Learning (AICML) and the Computer and Automation Research Institute of the Hungarian Academy of Sciences. This work was supported in part by the the Austrian Science Fund FWF (S9104-N04 SP4) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. We also acknowledge support by the PASCAL pump priming projects "Sequential Forecasting" and "Online Performance of Reinforcement Learning with Internal Reward Functions". This publication only reflects the authors' views.

References

1. Kleinberg, R.: Nearly tight bounds for the continuum-armed bandit problem. In: Advances in Neural Information Processing Systems 17 NIPS, 697–704 (2004)
2. Agrawal, R.: The continuum-armed bandit problem. SIAM J. Control Optim. 33, 1926–1951 (1995)
3. Cope, E.: Regret and convergence bounds for a class of continuum-armed bandit problems. submitted (2006)
4. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multi-armed bandit problem. Mach. Learn. 47, 235–256 (2002)
5. Cesa-Bianchi, N., Lugosi, G., Stoltz, G.: Minimizing regret with label efficient prediction. IEEE Trans. Inform. Theory 51, 2152–2162 (2004)