
Online Optimization in \mathcal{X} -Armed Bandits

Sébastien Bubeck

INRIA Lille - Nord Europe, SequeL project, France
sebastien.bubeck@inria.fr

Rémi Munos

INRIA Lille - Nord Europe, SequeL project, France
remi.munos@inria.fr

Gilles Stoltz

Ecole Normale Supérieure, CNRS, France and HEC Paris, CNRS, France
gilles.stoltz@ens.fr

Csaba Szepesvári

Department of Computing Science, University of Alberta Edmonton T6G 2E8, Canada
szepesva@cs.ualberta.ca

Abstract

We consider a generalization of stochastic bandit problems where the set of arms, \mathcal{X} , is allowed to be a generic topological space. We constraint the mean-payoff function with a dissimilarity function over \mathcal{X} in a way that is more general than Lipschitz. We construct an arm selection policy whose regret improves upon previous result for a large class of problems. In particular, our results imply that if \mathcal{X} is the unit hypercube in a Euclidean space and the mean-payoff function has a finite number of global maxima around which the behavior of the function is locally Hölder with a known exponent, then the expected regret is bounded up to a logarithmic factor by \sqrt{n} , i.e., the rate of the growth of the regret is independent of the dimension of the space. Moreover, we prove the minimax optimality of our algorithm for the class of mean-payoff functions we consider.

1 Introduction and motivation

Bandit problems arise in many settings, including clinical trials, scheduling, on-line parameter tuning of algorithms or optimization of controllers based on simulations. In the classical bandit problem there are a finite number of arms that the decision maker can select at discrete time steps. Selecting an arm results in a random reward, whose distribution is determined by the identity of the arm selected. The distributions associated with the arms are unknown to the decision maker whose goal is to maximize the expected sum of the rewards received.

In many practical situations the arms belong to a large set. This set could be continuous [1; 8; 4; 2; 9], hybrid-continuous, or it could be the space of infinite sequences over a finite alphabet [5]. In this paper we consider stochastic bandit problems where the set of arms, \mathcal{X} , is allowed to be an arbitrary topological space. We assume that the decision maker knows a dissimilarity function defined over this space that constraints the shape of the mean-payoff function. In particular, the dissimilarity function is assumed to put a lower bound on the mean-payoff function from below at each maxima. We also assume that the decision maker is able to cover the space of arms in a recursive manner, successively refining the regions in the covering such that the diameters of these sets shrink at a known geometric rate when measured with the dissimilarity.

Our work generalizes and improves previous works on continuum-armed bandit problems: Kleinberg [8] and Auer et al. [2] focussed on one-dimensional problems. Recently, Kleinberg et al. [9] considered generic metric spaces assuming that the mean-payoff function is Lipschitz with respect to the (known) metric of the space. They proposed an interesting algorithm that achieves essentially the best possible regret in a minimax sense with respect to these environments.

The goal of this paper is to further these works in a number of ways: (i) we allow the set of arms to be a generic topological space; (ii) we propose a practical algorithm motivated by the recent very successful tree-based optimization algorithms [10; 7; 5] and show that the algorithm is (iii) able to exploit higher order smoothness. In particular, as we shall argue in Section 7, (i) improves upon the results of Auer et al. [2], while (i), (ii) and (iii) improve upon the work of Kleinberg et al. [9]. Compared to Kleinberg et al. [9], our work represents an improvement in the fact that just like Auer et al. [2] we make use of the *local* properties of the mean-payoff function around the maxima only, and not a global property, such as Lipschitzness in the whole space. This allows us to obtain a regret which scales as $\tilde{O}(\sqrt{n})$ ¹ when e.g. the space is the unit hypercube and the mean-payoff function is locally Hölder with known exponent in the neighborhood of any maxima (which are in finite number) and bounded away from the maxima outside of these neighborhoods. Thus, we get the desirable property that the rate of growth of the regret is independent of the dimensionality of the input space. We also prove a minimax lower bound that matches our upper bound up to logarithmic factors, showing that the performance of our algorithm is essentially unimprovable in a minimax sense. Besides these theoretical advances the algorithm is anytime and easy to implement. Since it is based on ideas that have proved to be efficient, we expect it to perform well in practice and to make a significant impact on how on-line global optimization is performed.

2 Problem setup, notation

We consider a topological space \mathcal{X} , whose elements will be referred to as arms. A decision maker “pulls” the arms in \mathcal{X} one at a time at discrete time steps. Each pull results in a reward that depends on the arm chosen and which the decision maker learns of. The goal of the decision maker is to choose the arms so as to maximize the sum of the rewards that he receives. In this paper we are concerned with stochastic environments. Such an environment M associates to each arm $x \in \mathcal{X}$ a distribution M_x on the real line. The support of these distributions is assumed to be uniformly bounded with a known bound. For the sake of simplicity, we assume this bound is 1. We denote by $f(x)$ the expectation of M_x , which is assumed to be measurable (all measurability concepts are with respect to the Borel-algebra over \mathcal{X}). The function $f : \mathcal{X} \rightarrow \mathbb{R}$ thus defined is called the *mean-payoff function*. When in round n the decision maker pulls arm $X_n \in \mathcal{X}$, he receives a reward Y_n drawn from M_{X_n} , independently of the past arm choices and rewards.

A pulling strategy of a decision maker is determined by a sequence $\varphi = (\varphi_n)_{n \geq 1}$ of measurable mappings, where each φ_n maps the history space $\mathcal{H}_n = (\mathcal{X} \times [0, 1])^{n-1}$ to the space of probability measures over \mathcal{X} . By convention, φ_1 does not take any argument. A strategy is deterministic if for every n the range of φ_n contains only Dirac distributions.

¹We write $u_n = \tilde{O}(v_n)$ when $u_n = O(v_n)$ up to a logarithmic factor.

According to the process that was already informally described, a pulling strategy φ and an environment M jointly determine a random process $(X_1, Y_1, X_2, Y_2, \dots)$ in the following way: In round one, the decision maker draws an arm X_1 at random from φ_1 and gets a payoff Y_1 drawn from M_{X_1} . In round $n \geq 2$, first, X_n is drawn at random according to $\varphi_n(X_1, Y_1, \dots, X_{n-1}, Y_{n-1})$, but otherwise independently of the past. Then the decision maker gets a rewards Y_n drawn from M_{X_n} , independently of all other random variables in the past given X_n .

Let $f^* = \sup_{x \in \mathcal{X}} f(x)$ be the maximal expected payoff. The *cumulative regret* of a pulling strategy in environment M is $\widehat{R}_n = n f^* - \sum_{t=1}^n Y_t$, and the cumulative pseudo-regret is $R_n = n f^* - \sum_{t=1}^n f(X_t)$. In the sequel, we restrict our attention to the expected regret $\mathbb{E}[R_n]$, which in fact equals $\mathbb{E}[\widehat{R}_n]$, as can be seen by the application of the tower rule.

3 The Hierarchical Optimistic Optimization (HOO) strategy

3.1 Trees of coverings

We first introduce the notion of a tree of coverings. Our algorithm will require such a tree as an input.

Definition 1 (Tree of coverings). *A tree of coverings is a family of measurable subsets $(\mathcal{P}_{h,i})_{1 \leq i \leq 2^h, h \geq 0}$ of \mathcal{X} such that for all fixed integer $h \geq 0$, the covering $\cup_{1 \leq i \leq 2^h} \mathcal{P}_{h,i} = \mathcal{X}$ holds. Moreover, the elements of the covering are obtained recursively: each subset $\mathcal{P}_{h,i}$ is covered by the two subsets $\mathcal{P}_{h+1,2i-1}$ and $\mathcal{P}_{h+1,2i}$.*

A tree of coverings can be represented, as the name suggests, by a binary tree \mathcal{T} . The whole domain $\mathcal{X} = \mathcal{P}_{0,1}$ corresponds to the root of the tree and $\mathcal{P}_{h,i}$ corresponds to the i -th node of depth h , which will be referred to as node (h, i) in the sequel. The fact that each $\mathcal{P}_{h,i}$ is covered by the two subsets $\mathcal{P}_{h+1,2i-1}$ and $\mathcal{P}_{h+1,2i}$ corresponds to the childhood relationship in the tree. Although the definition allows the child-regions of a node to cover a larger part of the space, typically the size of the regions shrinks as depth h increases (cf. Assumption 1).

Remark 1. *Our algorithm will instantiate the nodes of the tree on an "as needed" basis, one by one. In fact, at any round n it will only need n nodes connected to the root.*

3.2 Statement of the HOO strategy

The algorithm picks at each round a node in the infinite tree \mathcal{T} as follows. In the first round, it chooses the root node $(0, 1)$. Now, consider round $n + 1$ with $n \geq 1$. Let us denote by \mathcal{T}_n the set of nodes that have been picked in previous rounds and by \mathcal{S}_n the nodes which are not in \mathcal{T}_n but whose parent is. The algorithm picks at round $n + 1$ a node $(H_{n+1}, I_{n+1}) \in \mathcal{S}_n$ according to the deterministic rule that will be described below. After selecting the node, the algorithm further chooses an arm $X_{n+1} \in \mathcal{P}_{H_{n+1}, I_{n+1}}$. This selection can be stochastic or deterministic. We do not put any further restriction on it. The algorithm then gets a reward Y_{n+1} as described above and the procedure goes on: (H_{n+1}, I_{n+1}) is added to \mathcal{T}_n to form \mathcal{T}_{n+1} and the children of (H_{n+1}, I_{n+1}) are added to \mathcal{S}_n to give rise to \mathcal{S}_{n+1} . Let us now turn to how (H_{n+1}, I_{n+1}) is selected.

Along with the nodes the algorithm stores what we call B -values. The node $(H_{n+1}, I_{n+1}) \in \mathcal{S}_n$ to expand at round $n + 1$ is picked by following a path from the root to a node in \mathcal{S}_n , where at each node along the path the child with the larger B -value is selected (ties are broken arbitrarily). In order to define a node's B -value, we need a few quantities. Let $\mathcal{C}(h, i)$ be the set that collects (h, i) and its descendants. We let

$$N_{h,i}(n) = \sum_{t=1}^n \mathbb{I}_{\{(H_t, I_t) \in \mathcal{C}(h,i)\}}$$

be the number of times the node (h, i) was visited. A given node (h, i) is always picked at most once, but since its descendants may be picked afterwards, subsequent paths in the tree can go through it. Consequently, $1 \leq N_{h,i}(n) \leq n$ for all nodes $(h, i) \in \mathcal{T}_n$. Let $\widehat{\mu}_{h,i}(n)$ be the empirical average of the rewards received for the time-points when the path followed by the algorithm went through (h, i) :

$$\widehat{\mu}_{h,i}(n) = \frac{1}{N_{h,i}(n)} \sum_{t=1}^n Y_t \mathbb{I}_{\{(H_t, I_t) \in \mathcal{C}(h,i)\}}.$$

The corresponding upper confidence bound is by definition

$$U_{h,i}(n) = \widehat{\mu}_{h,i}(n) + \sqrt{\frac{2 \ln n}{N_{h,i}(n)}} + \nu_1 \rho^h,$$

where $0 < \rho < 1$ and $\nu_1 > 0$ are parameters of the algorithm (to be chosen later by the decision maker, see Assumption 1). For nodes not in \mathcal{T}_n , by convention, $U_{h,i}(n) = +\infty$. Now, for a node (h, i) in \mathcal{S}_n , we define its B -value to be $B_{h,i}(n) = +\infty$. The B -values for nodes in \mathcal{T}_n are given by

$$B_{h,i}(n) = \min \left\{ U_{h,i}(n), \max \{ B_{h+1,2i-1}(n), B_{h+1,2i}(n) \} \right\}.$$

Note that the algorithm is deterministic (apart, maybe, from the arbitrary random choice of X_t in \mathcal{P}_{H_t, I_t}). Its total space requirement is linear in n while total running time at round n is at most quadratic in n , though we conjecture that it is $O(n \log n)$ on average.

4 Assumptions made on the model and statement of the main result

We suppose that \mathcal{X} is equipped with a *dissimilarity* ℓ , that is a non-negative mapping $\ell : \mathcal{X}^2 \rightarrow \mathbb{R}$ satisfying $\ell(x, x) = 0$. The diameter (with respect to ℓ) of a subset A of \mathcal{X} is given by $\text{diam } A = \sup_{x, y \in A} \ell(x, y)$. Given the dissimilarity ℓ , the ‘‘open’’ ball with radius $\varepsilon > 0$ and center $c \in \mathcal{X}$ is $\mathcal{B}(c, \varepsilon) = \{x \in \mathcal{X} : \ell(c, x) < \varepsilon\}$ (we do not require the topology induced by ℓ to be related to the topology of \mathcal{X} .) In what follows when we refer to an (open) ball, we refer to the ball defined with respect to ℓ . The dissimilarity will be used to capture the smoothness of the mean-payoff function. The decision maker chooses ℓ and the tree of coverings. The following assumption relates this choice to the parameters ρ and ν_1 of the algorithm:

Assumption 1. *There exist $\rho < 1$ and $\nu_1, \nu_2 > 0$ such that for all integers $h \geq 0$ and all $i = 1, \dots, 2^h$, the diameter of $\mathcal{P}_{h,i}$ is bounded by $\nu_1 \rho^h$, and $\mathcal{P}_{h,i}$ contains an open ball $\mathcal{P}'_{h,i}$ of radius $\nu_2 \rho^h$. For a given h , the $\mathcal{P}'_{h,i}$ are disjoint for $1 \leq i \leq 2^h$.*

Remark 2. *A typical choice for the coverings in a cubic domain is to let the domains be hyper-rectangles. They can be obtained, e.g., in a dyadic manner, by splitting at each step hyper-rectangles in the middle along their longest side, in an axis parallel manner; if all sides are equal, we split them along the first axis. In this example, if $\mathcal{X} = [0, 1]^D$ and $\ell(x, y) = \|x - y\|^\alpha$ then we can take $\rho = 2^{-\alpha/D}$, $\nu_1 = (\sqrt{D}/2)^\alpha$ and $\nu_2 = 1/8^\alpha$.*

The next assumption concerns the environment.

Definition 2. *We say that f is weakly Lipschitz with respect to ℓ if for all $x, y \in \mathcal{X}$,*

$$f^* - f(y) \leq f^* - f(x) + \max \{ f^* - f(x), \ell(x, y) \}. \quad (1)$$

Note that weak Lipschitzness is satisfied whenever f is 1-Lipschitz, i.e., for all $x, y \in \mathcal{X}$, one has $|f(x) - f(y)| \leq \ell(x, y)$. On the other hand, weak Lipschitzness implies local (one-sided) 1-Lipschitzness at any maxima. Indeed, at an optimal arm x^* (i.e., such that $f(x^*) = f^*$), (1) rewrites to $f(x^*) - f(y) \leq$

$\ell(x^*, y)$. However, weak Lipschitzness does not constraint the growth of the loss in the vicinity of other points. Further, weak Lipschitzness, unlike Lipschitzness, does not constraint the local *decrease* of the loss at any point. Thus, weak-Lipschitzness is a property that lies somewhere between a growth condition on the loss around optimal arms and (one-sided) Lipschitzness. Note that since weak Lipschitzness is defined with respect to a dissimilarity, it can actually capture higher-order smoothness at the optima. For example, $f(x) = 1 - x^2$ is weak Lipschitz with the dissimilarity $\ell(x, y) = c(x - y)^2$ for some appropriate constant c .

Assumption 2. *The mean-payoff function f is weakly Lipschitz.*

Let $f_{h,i}^* = \sup_{x \in \mathcal{P}_{h,i}} f(x)$ and $\Delta_{h,i} = f^* - f_{h,i}^*$ be the suboptimality of node (h, i) . We say that a node (h, i) is optimal (respectively, suboptimal) if $\Delta_{h,i} = 0$ (respectively, $\Delta_{h,i} > 0$). Let $\mathcal{X}_\varepsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X} : f(x) \geq f^* - \varepsilon\}$ be the set of ε -optimal arms. The following result follows from the definitions; a proof can be found in the appendix.

Lemma 1. *Let Assumption 1 and 2 hold. If the suboptimality $\Delta_{h,i}$ of a region is bounded by $c\nu_1\rho^h$ for some $c > 0$, then all arms in $\mathcal{P}_{h,i}$ are $\max\{2c, c + 1\}\nu_1\rho^h$ -optimal.*

The last assumption is closely related to Assumption 2 of Auer et al. [2], who observed that the regret of a continuum-armed bandit algorithm should depend on how fast the volume of the sets of ε -optimal arms shrinks as $\varepsilon \rightarrow 0$. Here, we capture this by defining a new notion, the near-optimality dimension of the mean-payoff function. The connection between these concepts, as well as the zooming dimension defined by Kleinberg et al. [9] will be further discussed in Section 7.

Define the packing number $\mathcal{P}(\mathcal{X}, \ell, \varepsilon)$ to be the size of the largest packing of \mathcal{X} with disjoint open balls of radius ε with respect to the dissimilarity ℓ .² We now define the near-optimality dimension, which characterizes the size of the sets \mathcal{X}_ε in terms of ε , and then state our main result.

Definition 3. *For $c > 0$ and $\varepsilon_0 > 0$, the (c, ε_0) -near-optimality dimension of f with respect to ℓ equals*

$$\inf \left\{ d \in [0, +\infty) : \exists C \text{ s.t. } \forall \varepsilon \leq \varepsilon_0, \mathcal{P}(\mathcal{X}_{c\varepsilon}, \ell, \varepsilon) \leq C \varepsilon^{-d} \right\} \quad (2)$$

(with the usual convention that $\inf \emptyset = +\infty$).

Theorem 1 (Main result). *Let Assumptions 1 and 2 hold and assume that the $(4\nu_1/\nu_2, \nu_2)$ -near-optimality dimension of the considered environment is $d < +\infty$. Then, for any $d' > d$ there exists a constant $C(d')$ such that for all $n \geq 1$,*

$$\mathbb{E}[R_n] \leq C(d') n^{(d'+1)/(d'+2)} (\ln n)^{1/(d'+2)}.$$

Further, if the near-optimality dimension is achieved, i.e., the infimum is achieved in (2), then the result holds also for $d' = d$.

Remark 3. *We can relax the weak-Lipschitz property by requiring it to hold only locally around the maxima. In fact, at the price of increased constants, the result continues to hold if there exists $\varepsilon > 0$ such that (1) holds for any $x, y \in \mathcal{X}_\varepsilon$. To show this we only need to carefully adapt the steps of the proof below. We omit the details from this extended abstract.*

5 Analysis of the regret and proof of the main result

We first state three lemmas, whose proofs can be found in the appendix. The proofs of Lemmas 3 and 4 rely on concentration-of-measure techniques, while that of Lemma 2 follows from a simple case study. Let us fix some path $(0, 1), (1, i_1^*), \dots, (h, i_h^*), \dots$, of optimal nodes, starting from the root.

²Note that sometimes packing numbers are defined as the largest packing with disjoint open balls of radius $\varepsilon/2$, or, ε -nets.

Lemma 2. Let (h, i) be a suboptimal node. Let k be the largest depth such that (k, i_k^*) is on the path from the root to (h, i) . Then we have

$$\mathbb{E}[N_{h,i}(n)] \leq u + \sum_{t=u+1}^n \mathbb{P}\left\{N_{h,i}(t) > u \text{ and } [U_{h,i}(t) > f^* \text{ or } U_{s,i_s^*} \leq f^* \text{ for some } s \in \{k+1, \dots, t-1\}]\right\}.$$

Lemma 3. Let Assumptions 1 and 2 hold. Then, for all optimal nodes and for all integers $n \geq 1$, $\mathbb{P}\{U_{h,i}(n) \leq f^*\} \leq n^{-3}$.

Lemma 4. Let Assumptions 1 and 2 hold. Then, for all integers $t \leq n$, for all suboptimal nodes (h, i) such that $\Delta_{h,i} > \nu_1 \rho^h$, and for all integers $u \geq 1$ such that $u \geq \frac{8 \ln n}{(\Delta_{h,i} - \nu_1 \rho^h)^2}$, one has $\mathbb{P}\{U_{h,i}(t) > f^* \text{ and } N_{h,i}(t) > u\} \leq t n^{-4}$.

Taking u as the integer part of $(8 \ln n)/(\Delta_{h,i} - \nu_1 \rho^h)^2$, and combining the results of Lemma 2, 3, and 4 with a union bound leads to the following key result.

Lemma 5. Under Assumptions 1 and 2, for all suboptimal nodes (h, i) such that $\Delta_{h,i} > \nu_1 \rho^h$, we have, for all $n \geq 1$,

$$\mathbb{E}[N_{h,i}(n)] \leq \frac{8 \ln n}{(\Delta_{h,i} - \nu_1 \rho^h)^2} + \frac{2}{n}.$$

We are now ready to prove Theorem 1.

Proof. For the sake of simplicity we assume that the infimum in the definition of near-optimality is achieved. To obtain the result in the general case one only needs to replace d below by $d' > d$ in the proof below.

First step. For all $h = 1, 2, \dots$, denote by \mathcal{I}_h the nodes at depth h that are $2\nu_1 \rho^h$ -optimal, i.e., the nodes (h, i) such that $f_{h,i}^* \geq f^* - 2\nu_1 \rho^h$. Then, \mathcal{I} is the union of these sets of nodes. Further, let \mathcal{J} be the set of nodes that are not in \mathcal{I} but whose parent is in \mathcal{I} . We then denote by \mathcal{J}_h the nodes in \mathcal{J} that are located at depth h in the tree. Lemma 4 bounds the expected number of times each node $(h, i) \in \mathcal{J}_h$ is visited. Since $\Delta_{h,i} > 2\nu_1 \rho^h$, we get

$$\mathbb{E}[N_{h,i}(n)] \leq \frac{8 \ln n}{\nu_1^2 \rho^{2h}} + \frac{2}{n}.$$

Second step. We bound here the cardinality $|\mathcal{I}_h|$, $h > 0$. If $(h, i) \in \mathcal{I}_h$ then since $\Delta_{h,i} \leq 2\nu_1 \rho^h$, by Lemma 1 $\mathcal{P}_{h,i} \subset \mathcal{X}_{4\nu_1 \rho^h}$. Since by Assumption 1, the sets $(\mathcal{P}_{h,i})$, for $(h, i) \in \mathcal{I}_h$, contain disjoint balls of radius $\nu_2 \rho^h$, we have that

$$|\mathcal{I}_h| \leq \mathcal{P}(\cup_{(h,i) \in \mathcal{I}_h} \mathcal{P}_{h,i}, \ell, \nu_2 \rho^h) \leq \mathcal{P}(\mathcal{X}_{(4\nu_1/\nu_2) \nu_2 \rho^h}, \ell, \nu_2 \rho^h) \leq C (\nu_2 \rho^h)^{-d},$$

where we used the assumption that d is the $(4\nu_1/\nu_2, \nu_2)$ -near-optimality dimension of f (and C is the constant introduced in the definition of the near-optimality dimension).

Third step. Choose $\eta > 0$ and let H be the smallest integer such that $\rho^H \leq \eta$. We partition the infinite tree \mathcal{T} into three sets of nodes, $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{T}_3$. The set \mathcal{T}_1 contains nodes of \mathcal{I}_H and their descendants, $\mathcal{T}_2 = \cup_{0 \leq h < H} \mathcal{I}_h$, and \mathcal{T}_3 contains the nodes $\cup_{1 \leq h \leq H} \mathcal{J}_h$ and their descendants. (Note that \mathcal{T}_1 and \mathcal{T}_3 are potentially infinite, while \mathcal{T}_2 is finite.)

We denote by (H_t, I_t) the node that was chosen by the forecaster at round t to pick X_t . From the definition of the forecaster, no two such random variables are equal, since each node is picked at most once. We decompose the regret according to the element \mathcal{T}_j where the chosen nodes (H_t, I_t) belong to:

$$\mathbb{E}[R_n] = \mathbb{E}\left[\sum_{t=1}^n (f^* - f(X_t))\right] = \mathbb{E}[R_{n,1}] + \mathbb{E}[R_{n,2}] + \mathbb{E}[R_{n,3}],$$

$$\text{where for all } i = 1, 2, 3, \quad R_{n,i} = \sum_{t=1}^n (f^* - f(X_t)) \mathbb{1}_{\{(H_t, I_t) \in \mathcal{T}_i\}}.$$

The contribution from \mathcal{T}_1 is easy to bound. By definition any node in \mathcal{I}_H is $2\nu_1\rho^H$ -optimal. Hence, by Lemma 1 the corresponding domain is included in $\mathcal{X}_{4\nu_1\rho^H}$. The domains of these nodes' descendants are of course still included in $\mathcal{X}_{4\nu_1\rho^H}$. Therefore, $\mathbb{E}[R_{n,1}] \leq 4n\nu_1\rho^H$.

For $h \geq 1$, consider a node $(h, i) \in \mathcal{T}_2$. It belongs to \mathcal{I}_h and is therefore $2\nu_1\rho^h$ -optimal. By Lemma 1, the corresponding domain is included in $\mathcal{X}_{4\nu_1\rho^h}$. By the result of the second step and using that each node is played at most once, one gets

$$\mathbb{E}[R_{n,2}] \leq \sum_{h=0}^{H-1} 4\nu_1\rho^h |\mathcal{I}_h| \leq 4\nu_1 C \nu_2^{-d} \sum_{h=0}^{H-1} \rho^{h(1-d)}.$$

We finish with the contribution from \mathcal{T}_3 . We first remark that since the parent of any element $(h, i) \in \mathcal{J}_h$ is in \mathcal{I}_{h-1} , by Lemma 1 again, we have that $\mathcal{P}_{h,i} \subset \mathcal{X}_{4\nu_1\rho^{h-1}}$. To each node (H_t, I_t) played in \mathcal{T}_3 , we associate the element (H'_t, I'_t) of some \mathcal{J}_h on the path from the root to (H_t, I_t) . When (H_t, I_t) is played, the chosen arm X_t belongs also to $\mathcal{P}_{H'_t, I'_t}$. Decomposing $R_{n,3}$ according to the elements of $\cup_{1 \leq h \leq H} \mathcal{J}_h$, we then bound the regret from \mathcal{T}_3 as

$$\mathbb{E}[R_{n,3}] \leq \sum_{h=1}^H 4\nu_1\rho^{h-1} \sum_{i: (h,i) \in \mathcal{J}_h} \mathbb{E}[N_{h,i}(n)] \leq \sum_{h=1}^H 4\nu_1\rho^{h-1} |\mathcal{J}_h| \left(\frac{8 \ln n}{\nu_1^2 \rho^{2h}} + \frac{2}{n} \right)$$

where we used the result of the first step. Now, it follows from that fact that the parent of \mathcal{J}_h is in \mathcal{I}_{h-1} that $|\mathcal{J}_h| \leq 2|\mathcal{I}_{h-1}|$. Substituting this and the bound on $|\mathcal{I}_{h-1}|$, we get

$$\mathbb{E}[R_{n,3}] \leq 8\nu_1 C \nu_2^{-d} \sum_{h=1}^H \rho^{h(1-d)+d-1} \left(\frac{8 \ln n}{\nu_1^2 \rho^{2h}} + \frac{2}{n} \right).$$

Fourth step. Putting things together, we have proved

$$\begin{aligned} \mathbb{E}[R_n] &\leq 4n\nu_1\rho^H + 4\nu_1 C \nu_2^{-d} \sum_{h=0}^{H-1} \rho^{h(1-d)} + 8\nu_1 C \nu_2^{-d} \sum_{h=1}^H \rho^{h(1-d)+d-1} \left(\frac{8 \ln n}{\nu_1^2 \rho^{2h}} + \frac{2}{n} \right) \\ &= O\left(n\rho^H + (\ln n) \sum_{h=1}^H \rho^{-h(1+d)}\right) = O\left(n\rho^H + \rho^{-H(1+d)} \ln n\right) = O\left(n^{(d+1)/(d+2)} (\ln n)^{1/(d+2)}\right) \end{aligned}$$

by using first that $\rho < 1$ and then, by optimizing over ρ^H (the worst value being $\rho^H \sim (\frac{n}{\ln n})^{-1/(d+2)}$). \square

6 Minimax optimality

The packing dimension of a set \mathcal{X} is the smallest d such that there exists a constant k such that for all $\varepsilon > 0$, $\mathcal{P}(\mathcal{X}, \ell, \varepsilon) \leq k\varepsilon^{-d}$. For instance, compact subsets of \mathbb{R}^d (with non-empty interior) have a packing dimension of d whenever ℓ is a norm. If \mathcal{X} has a packing dimension of d , then all environments have a near-optimality dimension less than d . The proof of the main theorem indicates that the constant $C(d)$ only depends on d, k (of the definition of packing dimension), ν_1, ν_2 , and ρ , but not on the environment as long as it is weakly Lipschitz. Hence, we can extract from it a distribution-free bound of the form $\tilde{O}(n^{(d+1)/(d+2)})$. In fact, this bound can be shown to be optimal as is illustrated by the theorem below, whose assumptions are satisfied by, e.g., compact subsets of \mathbb{R}^d and if ℓ is some norm of \mathbb{R}^d . The proof can be found in the appendix.

Theorem 2. *If \mathcal{X} is such that there exists $c > 0$ with $\mathcal{P}(\mathcal{X}, \ell, \varepsilon) \geq c\varepsilon^{-d} \geq 2$ for all $\varepsilon \leq 1/4$ then for all $n \geq 4^{d-1} c / \ln(4/3)$, all strategies φ are bound to suffer a regret of at least*

$$\sup \mathbb{E} R_n(\varphi) \geq \frac{1}{4} \left(\frac{1}{4} \sqrt{\frac{c}{4 \ln(4/3)}} \right)^{2/(d+2)} n^{(d+1)/(d+2)},$$

where the supremum is taken over all environments with weakly Lipschitz payoff functions.

7 Discussion

Several works [1; 8; 4; 2; 9] have considered continuum-armed bandits in Euclidean or metric spaces and provided upper- and lower-bounds on the regret for given classes of environments. Cope [4] derived a regret of $\tilde{O}(\sqrt{n})$ for compact and convex subset of R^d and a mean-payoff function with unique minima and second order smoothness. Kleinberg [8] considered mean-payoff functions f on the real line that are Hölder with degree $0 < \alpha \leq 1$. The derived regret is $\Theta(n^{(\alpha+1)/(\alpha+2)})$. Auer et al. [2] extended the analysis to classes of functions with only a local Hölder assumption around maximum (with possibly higher smoothness degree $\alpha \in [0, \infty)$), and derived the regret $\Theta(n^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}})$, where β is such that the Lebesgue measure of ε -optimal states is $O(\varepsilon^\beta)$. Another setting is that of [9] who considered a metric space (\mathcal{X}, ℓ) and assumed that f is Lipschitz w.r.t. ℓ . The obtained regret is $\tilde{O}(n^{(d+1)/(d+2)})$ where d is the zooming dimension (defined similarly to our near-optimality dimension, but using covering numbers instead of packing numbers and the sets $\mathcal{X}_\varepsilon \setminus \mathcal{X}_{\varepsilon/2}$). When (\mathcal{X}, ℓ) is a metric space covering and packing numbers are equivalent and we may prove that the zooming dimension and near-optimality dimensions are equal.

Our main contribution compared to [9] is that our weak-Lipschitz assumption, which is substantially weaker than the global Lipschitz assumption assumed in [9], enables our algorithm to work better in some common situations, such as when the mean-payoff function assumes a local smoothness whose order is larger than one. In order to relate all these results, let us consider a specific example: Let $\mathcal{X} = [0, 1]^D$ and assume that the mean-reward function f is locally equivalent to a Hölder function with degree $\alpha \in [0, \infty)$ around any maxima x^* of f (the number of maxima is assumed to be finite):

$$f(x^*) - f(x) = \Theta(\|x - x^*\|^\alpha) \text{ as } x \rightarrow x^*. \quad (3)$$

This means that $\exists c_1, c_2, \varepsilon_0 > 0, \forall x$, s.t. $\|x - x^*\| \leq \varepsilon_0, c_1\|x - x^*\|^\alpha \leq f(x^*) - f(x) \leq c_2\|x - x^*\|^\alpha$. Under this assumption, the result of Auer et al. [2] shows that for $D = 1$, the regret is $\Theta(\sqrt{n})$ (since here $\beta = 1/\alpha$). Our result allows us to extend the \sqrt{n} regret rate to any dimension D . Indeed, if we choose our dissimilarity measure to be $\ell_\alpha(x, y) \stackrel{\text{def}}{=} \|x - y\|^\alpha$, we may prove that f satisfies a locally weak-Lipschitz condition (as defined in Remark 3) and that the near-optimality dimension is 0. Thus our regret is $\tilde{O}(\sqrt{n})$, i.e., the rate is independent of the dimension D .

In comparison, since Kleinberg et al. [9] have to satisfy a global Lipschitz assumption, they can not use ℓ_α when $\alpha > 1$. Indeed a function globally Lipschitz with respect to ℓ_α is essentially constant. Moreover ℓ_α does not define a metric for $\alpha > 1$. If one resort to the Euclidean metric to fulfill their requirement that f be Lipschitz w.r.t. the metric then the zooming dimension becomes $D(\alpha - 1)/\alpha$, while the regret becomes $\tilde{O}(n^{(D(\alpha-1)+\alpha)/(D(\alpha-1)+2\alpha)})$, which is strictly worse than $\tilde{O}(\sqrt{n})$ and in fact becomes close to the slow rate $\tilde{O}(n^{(D+1)/(D+2)})$ when α is larger. Nevertheless, in the case of $\alpha \leq 1$ they get the same regret rate.

In contrast, our result shows that under very weak constraints on the mean-payoff function and if the local behavior of the function around its maximum (or finite number of maxima) is known then global optimization suffers a regret of order $\tilde{O}(\sqrt{n})$, independent of the space dimension. As an interesting sidenote let us also remark that our results allow different smoothness orders along different dimensions, i.e., heterogenous smoothness spaces.

A Proof of Lemma 1

Proof. We denote by $x_{h,i}^*(\delta)$ an element of $\mathcal{P}_{h,i}$ such that

$$f(x_{h,i}^*(\delta)) \geq f_{h,i}^* - \delta.$$

By the weakly Lipschitz property, it then follows that for all $y \in \mathcal{P}_{h,i}$,

$$f^* - f(y) \leq f^* - f(x_{h,i}^*(\delta)) + \max\{f^* - f(x_{h,i}^*(\delta)), \ell(x_{h,i}^*(\delta), y)\} \leq \Delta_{h,i} + \delta + \max\{\Delta_{h,i} + \delta, \text{diam } \mathcal{P}_{h,i}\}.$$

Letting $\delta \rightarrow 0$ and substituting the bounds on the suboptimality and on the diameter of $\mathcal{P}_{h,i}$ concludes the proof. \square

B Proof of Lemma 2

Proof. We consider a given round $t \in \{1, \dots, n\}$. If $(H_t, I_t) \in \mathcal{C}(h, i)$, then this is because the child of (k, i_k^*) on the path to (h, i) had a better B -value than its brother $(k+1, i_{k+1}^*)$. Since by definition, B -values can only decrease on a path, this entails that $B_{h,i}(t) \geq B_{k+1, i_{k+1}^*}(t)$. This in turn implies, again by definition of the B -values, that $U_{h,i}(t) \geq B_{k+1, i_{k+1}^*}(t)$. Thus,

$$\{(H_t, I_t) \in \mathcal{C}(h, i)\} \subset \{U_{h,i}(t) \geq B_{k+1, i_{k+1}^*}(t)\} \subset \{U_{h,i}(t) \geq f^*\} \cup \{B_{k+1, i_{k+1}^*}(t) \leq f^*\}.$$

But, once again by definition of B -values,

$$\{B_{k+1, i_{k+1}^*}(t) \leq f^*\} \subset \{U_{k+1, i_{k+1}^*}(t) \leq f^*\} \cup \{B_{k+2, i_{k+2}^*}(t) \leq f^*\},$$

and the argument can be iterated. Since at round t not more than t nodes have been played (including the suboptimal (h, i)), we know that (t, i_t^*) and its descendants have U -values and B -values equal to $+\infty$. We thus have proved the inclusion

$$\{(H_t, I_t) \in \mathcal{C}(h, i)\} \subset \{U_{h,i}(t) \geq f^*\} \cup \left(\{B_{k+1, i_{k+1}^*}(t) \leq f^*\} \cup \dots \cup \{B_{t-1, i_{t-1}^*}(t) \leq f^*\} \right).$$

The result follows by simply distinguishing whether $N_{h,i}(t) > u$ (which can only happen if $t \geq u$) or not. \square

C Proof of Lemma 3

Proof. $U_{h,i} \leq f^*$ is not true when node (h, i) was never pulled (in this case, by definition, $U_{h,i}(n) = +\infty$). We may thus conduct the study in the sequel on the event $\{N_{h,i}(n) \geq 1\}$.

Lemma 1 with $c = 0$ gives that $f^* - f(x) \leq \nu_1 \rho^h$ holds for any arm $x \in \mathcal{P}_{h,i}$. Hence,

$$\sum_{t=1}^n (f(X_t) + \nu_1 \rho^h - f^*) \mathbb{1}_{\{(H_t, I_t) \in \mathcal{C}(h, i)\}} \geq 0$$

and therefore,

$$\begin{aligned} & \mathbb{P}\{U_{h,i}(n) \leq f^* \text{ and } N_{h,i}(n) \geq 1\} \\ &= \mathbb{P}\left\{\widehat{\mu}_{h,i}(n) + \sqrt{\frac{2 \ln n}{N_{h,i}(n)}} + \nu_1 \rho^h \leq f^* \text{ and } N_{h,i}(n) \geq 1\right\} \\ &= \mathbb{P}\left\{N_{h,i}(n) \widehat{\mu}_{h,i}(n) + N_{h,i}(n) (\nu_1 \rho^h - f^*) \leq -\sqrt{N_{h,i}(n) 2 \ln n} \text{ and } N_{h,i}(n) \geq 1\right\} \\ &\leq \mathbb{P}\left\{\sum_{t=1}^n (f(X_t) - Y_t) \mathbb{1}_{\{(H_t, I_t) \in \mathcal{C}(h, i)\}} \geq \sqrt{N_{h,i}(n) 2 \ln n} \text{ and } N_{h,i}(n) \geq 1\right\}. \end{aligned}$$

We take care of the last term with a union bound and the Hoeffding-Azuma inequality for martingale differences. To do this properly we need to define a sequence of (random) times when arms in $\mathcal{C}(h, i)$ were pulled:

$$T_j = \min \{t : N_{h,i}(t) = j\}, \quad j = 1, 2, \dots$$

Note that $1 \leq T_1 < T_2 < \dots$ and hence it holds that $T_j \geq j$. With these notation, $\tilde{X}_j = X_{T_j}$ is the j -th arm pulled in a domain corresponding to $\mathcal{C}(h, i)$, $\tilde{Y}_j = Y_{T_j}$ is the corresponding reward, and

$$\begin{aligned} & \mathbb{P} \left\{ \sum_{t=1}^n (f(X_t) - Y_t) \mathbb{1}_{\{(H_t, I_t) \in \mathcal{C}(h, i)\}} \geq \sqrt{N_{h,i}(n) 2 \ln n} \text{ and } N_{h,i}(n) \geq 1 \right\} \\ &= \mathbb{P} \left\{ \sum_{j=1}^{N_{h,i}(n)} (f(\tilde{X}_j) - \tilde{Y}_j) \geq \sqrt{N_{h,i}(n) 2 \ln n} \text{ and } N_{h,i}(n) \geq 1 \right\} \\ &\leq \sum_{t=1}^n \mathbb{P} \left\{ \sum_{j=1}^t (f(\tilde{X}_j) - \tilde{Y}_j) \geq \sqrt{2 t \ln n} \right\} \end{aligned}$$

where we used a union bound to get the last inequality.

We now prove that

$$Z_t = \sum_{j=1}^t (f(\tilde{X}_j) - \tilde{Y}_j)$$

is a martingale difference sequence (with respect to the filtration it generates). This follows, via optional skipping (see [6], Theorem 2.3), from the fact that

$$\sum_{t=1}^n (f(X_t) - Y_t) \mathbb{1}_{\{(H_t, I_t) \in \mathcal{C}(h, i)\}}$$

is a martingale, with respect to the filtration $\mathcal{F}_t = \sigma(X_1, Y_1, \dots, X_t, Y_t)$, and that $\{T_j = k\} \in \mathcal{F}_{k-1}$.

Applying the Hoeffding-Azuma inequality (using the bounded ranges), we then get, for each $t \geq 1$,

$$\mathbb{P} \left\{ \sum_{j=1}^t (f(\tilde{X}_j) - \tilde{Y}_j) \geq \sqrt{2 t \ln n} \right\} \leq \exp \left(-\frac{2 (\sqrt{2 t \ln n})^2}{t} \right) = n^{-4},$$

which concludes the proof. □

D Proof of Lemma 4

Proof. Remark that for the u mentioned in the statement of the lemma,

$$\sqrt{\frac{2 \ln t}{u}} + \nu_1 \rho^h \leq (\Delta_{h,i} + \nu_1 \rho^h) / 2,$$

and therefore,

$$\begin{aligned}
& \mathbb{P}\{U_{h,i}(t) > f^* \text{ and } N_{h,i}(t) > u\} \\
&= \mathbb{P}\left\{\widehat{\mu}_{h,i}(t) + \sqrt{\frac{2 \ln t}{N_{h,i}(t)}} + \nu_1 \rho^h > f_{h,i}^* + \Delta_{h,i} \text{ and } N_{h,i}(t) > u\right\} \\
&\leq \mathbb{P}\left\{\widehat{\mu}_{h,i}(t) > f_{h,i}^* + \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} \text{ and } N_{h,i}(t) > u\right\} \\
&\leq \mathbb{P}\left\{N_{h,i}(t) (\widehat{\mu}_{h,i}(t) - f_{h,i}^*) > \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} u \text{ and } N_{h,i}(t) > u\right\} \\
&= \mathbb{P}\left\{\sum_{s=1}^t (Y_s - f_{h,i}^*) \mathbb{I}_{\{(H_s, I_s) \in \mathcal{C}(h,i)\}} > \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} u \text{ and } N_{h,i}(t) > u\right\} \\
&\leq \mathbb{P}\left\{\sum_{s=1}^t (Y_s - f(X_s)) \mathbb{I}_{\{(H_s, I_s) \in \mathcal{C}(h,i)\}} > \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} u \text{ and } N_{h,i}(t) > u\right\}.
\end{aligned}$$

Now it follows again by the optional skipping argument, the Hoeffding-Azuma inequality, and a union bound, that

$$\begin{aligned}
& \mathbb{P}\left\{\sum_{s=1}^t (Y_s - f(X_s)) \mathbb{I}_{\{(H_s, I_s) \in \mathcal{C}(h,i)\}} > \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} u \text{ and } N_{h,i}(t) > u\right\} \\
&\leq \sum_{s=u+1}^t \exp\left(-\frac{2}{s} \left(\frac{(\Delta_{h,i} - \nu_1 \rho^h) u}{2}\right)^2\right) \leq t \exp\left(-\frac{1}{2} u (\Delta_{h,i} - \nu_1 \rho^h)^2\right) \leq t n^{-4}
\end{aligned}$$

(where we used the stated bound on u to obtain the last inequality). \square

E Proof of Theorem 2

We only deal with the case of deterministic strategies. The extension to randomized strategies can be done using Fubini's theorem.

For $\eta \in [0, 1/4]$ and $x^* \in \mathcal{X}$, we denote by f_{η, x^*} the mapping defined by

$$f_{\eta, x^*}(x) = \max\{\eta - \ell(x, x^*), 0\}$$

for all $x \in \mathcal{X}$ and by M_{η, x^*} the environment defined by

$$M_{\eta, x^*}(x) = \text{Ber}\left(\frac{1}{2} + f_{\eta, x^*}(x)\right)$$

for all $x \in \mathcal{X}$. We consider K points x_1, \dots, x_K in \mathcal{X} such that the balls $B_{x_j, \eta}$ with radius η centered at each of the x_j are non-overlapping. Note that $B_{x_j, \eta}$ is the support of f_{η, x^*} . In addition, the mean functions of all the defined environments are 1-Lipschitz and thus are weakly Lipschitz.

We will also need to consider environments on a finite set of arms $\{1, \dots, K+1\}$. We construct K different product-distributions $\nu_1, \nu_2, \dots, \nu_K$ for the arms $\{1, \dots, K+1\}$ as follows. For a given ν_j , the reward distribution associated to the i -th arm is $\nu_{j,i} = \text{Ber}(1/2)$ for all $i \neq j$ and $\nu_{j,j} = \text{Ber}(1/2 + \eta)$.

To each (deterministic) strategy φ on \mathcal{X} , we associate a random strategy ψ on the finite set of arms $\{1, \dots, K+1\}$ as follows. Let $t \geq 1$. Since φ is deterministic it associates to each sequence of rewards

$\{r_1, \dots, r_{t-1}\} \in \{0, 1\}^{t-1}$ a unique sequence $\{x_1, \dots, x_t\} \in \mathcal{X}^t$ of arms that φ would have pulled under this sequence of rewards. With a slight abuse of notation we can write $\varphi(r_1, \dots, r_{t-1}) = (x_1, \dots, x_t)$. Now assume that the historic of ψ at time t is $X_1, R_1, \dots, X_{t-1}, R_{t-1}$ and let $(X'_1, \dots, X'_t) = \varphi(R_1, \dots, R_{t-1})$. We then define

$$\begin{aligned} \psi_t &= \delta_{K+1} && \text{if } X'_t \notin \cup_j B_{x_j, \eta}, \\ \psi_t &= \left(1 - \frac{\ell(X'_t, x_j)}{\eta}\right) \delta_{x_j} + \frac{\ell(X'_t, x_j)}{\eta} \delta_{K+1} && \text{if } X'_t \in B_{x_j, \eta}, \end{aligned}$$

where δ_j is a dirac distribution on j .

We now want to prove that the distributions of the regrets for φ under M_{η, x_j} and for ψ under ν_j are equal for all $j = 1, \dots, K$. On the one hand, the expectations of the best arms are $1/2 + \eta$ under all these environments. On the other hand we can prove recursively that for any $\{r_1, \dots, r_t\} \in \{0, 1\}^t$,

$$\mathbb{P}(R_1 = r_1, \dots, R_t = r_t) = \mathbb{P}(R'_1 = r_1, \dots, R'_t = r_t).$$

where R_1, \dots, R_t (respectively R'_1, \dots, R'_t) is the sequence of rewards obtained by φ under M_{η, x_j} (respectively ψ under ν_j). The result is easy to check for $t = 1$ and for $t > 1$ it follows from

$$\mathbb{P}(R_1 = r_1, \dots, R_t = r_t) = \mathbb{P}(R_t = r_t | R_1 = r_1, \dots, R_{t-1} = r_{t-1}) \mathbb{P}(R_1 = r_1, \dots, R_{t-1} = r_{t-1})$$

and the same calculation for R'_t .

As a consequence, the regrets $R_n(\varphi)$ and $R_n(\psi)$ have the same expectation, that is, for all $j = 1, \dots, K$,

$$\mathbb{E}_j R_n(\varphi) = \mathbb{E}'_j R_n(\psi) \quad (4)$$

where \mathbb{E}_j denotes the expectation under M_{η, x_j} and \mathbb{E}'_j the one under ν_j .

But it can be extracted from the proof of the lower bound of [3, Section 6.9] that for all strategies ψ' , all $\eta \in [0, 1/4]$, and all integers K ,

$$\max_{j=1, \dots, K} \mathbb{E}'_j R_n(\psi') \geq \eta n \left(1 - \frac{1}{K} - \eta \sqrt{4 \ln(4/3) \frac{n}{K}}\right). \quad (5)$$

By the assumption on packing dimension, there exists $c > 0$ such that $K = c \eta^{-d} \geq 2$ is a suitable choice. Substituting this value, we get

$$\max_{j=1, \dots, K} \mathbb{E}_j R_n(\varphi) = \max_{j=1, \dots, K} \mathbb{E}'_j R_n(\psi) \geq \eta n \left(\frac{1}{2} - \eta^{1+d/2} \sqrt{\frac{4 \ln(4/3)}{c} n}\right).$$

The left-hand side is smaller than the maximal regret with respect to all weak-Lipschitz environments; the right-hand side can be optimized over $\eta \leq 1/4$ to get the claimed bound, by taking

$$\eta = \left(\frac{1}{4} \sqrt{\frac{c}{4 \ln(4/3)}}\right)^{2/(d+2)} n^{-1/(d+2)}.$$

References

- [1] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.
- [2] P. Auer, R. Ortner, and Cs. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. *20th Conference on Learning Theory*, pages 454–468, 2007.
- [3] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, UK, 2006.

- [4] E. Cope. Regret and convergence bounds for immediate-reward reinforcement learning with continuous action spaces. Preprint, 2004.
- [5] P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Proceedings of 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.
- [6] J.L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.
- [7] S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in Monte-Carlo go. Technical Report RR-6062, INRIA, 2006.
- [8] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems*, 2004.
- [9] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.
- [10] L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.