Lipschitz Bandits without the Lipschitz Constant

Sébastien Bubeck, Gilles Stoltz, Jia Yuan Yu

Lipschitz Bandits without the Lipschitz Constant

Sébastien Bubeck[1], Gilles Stoltz[2,3], and Jia Yuan Yu[2,3]

[1] Centre de Recerca Matemàtica, Barcelona, Spain
[2] Ecole normale supérieure, CNRS, Paris, France
[3] HEC Paris, CNRS, Jouy-en-Josas, France

Abstract. We consider the setting of stochastic bandit problems with a continuum of arms. We first point out that the strategies considered so far in the literature only provided theoretical guarantees of the form: given some tuning parameters, the regret is small with respect to a class of environments that depends on these parameters. This is however not the right perspective, as it is the strategy that should adapt to the specific bandit environment at hand, and not the other way round. Put differently, an adaptation issue is raised. We solve it for the special case of environments whose mean-payoff functions are globally Lipschitz. More precisely, we show that the minimax optimal orders of magnitude $L^{d/(d+2)}\,T^{(d+1)/(d+2)}$ of the regret bound against an environment f with Lipschitz constant L over T time instances can be achieved without knowing L or T in advance. This is in contrast to all previously known strategies, which require to some extent the knowledge of L to achieve this performance guarantee.

1 Introduction

In the (stochastic) bandit problem, a gambler tries to maximize the revenue gained by sequentially playing one of a finite number of arms that are each associated with initially unknown (and potentially different) payoff distributions [Rob52]. The gambler selects and pulls arms one by one in a sequential manner, simultaneously learning about the machines' payoff-distributions and accumulating rewards (or losses). Thus, in order to maximize his gain, the gambler must choose the next arm by taking into consideration both the urgency of gaining reward ("exploitation") and acquiring new information ("exploration"). Maximizing the total cumulative payoff is equivalent to minimizing the (total) *regret*, that is, minimizing the difference between the total cumulative payoff of the gambler and that of another clairvoyant gambler who chooses the arm with the best mean-payoff in every round. The quality of the gambler's strategy can be characterized by the rate of growth of his expected regret with time. In particular, if this rate of growth is sublinear, the gambler in the long run plays as well as his clairvoyant counterpart.

Continuum-armed bandit problems. Although the early papers studied bandits with a finite number of arms, researchers soon realized that bandits with infinitely many arms are also interesting, as well as practically significant. One

particularly important case is when the arms are identified by a finite number of continuous-valued parameters, resulting in online optimization problems over continuous finite-dimensional spaces. During the last decades numerous contributions have investigated such continuum-armed bandit problems, starting from the early formulations of [Agr95,Cop09,Kle04] to the more recent approaches of [AOS07,KSU08,BMSS11]. A special case of interest, which forms a bridge between the case of a finite number of arms and the continuum-armed setting, is the problem of bandit linear optimization, see [DHK08] and the references therein.

Not the right perspective! We call an environment f the mapping that associates with each arm $x \in \mathcal{X}$ the expectation $f(x)$ of its associated probability distribution. The theoretical guarantees given in the literature mentioned above are of the form: given some tuning parameters, the strategy is competitive, and sometimes even minimax optimal, with respect to a large class of environments that unfortunately depends on these parameters. But of course, this is not the right perspective: it is the strategy that should adapt to the environment, not the other way round!

More precisely, these parameters describe the smoothness of the environments f in the class at hand in terms of a global regularity and/or local regularities around the global maxima of f. The issues raised by some of the works mentioned above can be roughly described as follows:

- The class of environments for the CAB1 algorithm of [Kle04] is formed by environments that are (α, L, δ)–uniformly locally Lipschitz and the strategy CAB1 needs to know α to get the optimal dependency in the number T of arms pulled;
- For the Zooming algorithm of [KSU08], it is formed by environments that are 1–Lipschitz with respect to a fixed and known metric L;
- The HOO algorithm of [BMSS11] basically needs to know the pseudo-metric ℓ with respect to which f is weakly Lipschitz continuous, with Lipschitz constant equal to 1;
- Other examples include the UCB-air algorithm (which relies on a smoothness parameter β, see [WAM09]), the OLOP algorithm (smoothness parameter γ, see [BM10]), the LSE algorithm (smoothness parameter C_L, see [YM11]), and so on.

Adaptation to the unknown smoothness is needed. In a nutshell, adaptive methods are required. By adaptive methods, we mean—as is done in the statistical literature—agnostic methods, i.e., with minimal prior knowledge about f, that nonetheless obtain almost the same performance against a given environment f as if its smoothness were known beforehand.

More precisely, given a fixed (possibly vector-valued) parameter L lying in a set $\mathcal{L}$ and a class of allowed environments $\mathcal{F}_L$, where $L \in \mathcal{L}$, existing works present algorithms that are such that their worst-case regret bound over T time steps against environments in $\mathcal{F}_L$,

$$\sup_{f \in \mathcal{F}_L} R_T(f) \leqslant \varphi(T, L)\,,$$

is small and even minimax optimal, i.e., such that it has the optimal dependencies on T and L. However, to do so, the knowledge of L is required. In this work, we are given a much larger class of environments $\mathcal{F} = \cup_{L \in \mathcal{L}} \mathcal{F}_L$, and our goal is an algorithm that adapts in finite time to every instance f of $\mathcal{F}$, in the sense that for all T and $f \in \mathcal{F}$, the regret $R_T(f)$ is at most of the order of $\min \varphi(T, L)$, where the minimum is over the parameters L such that $f \in \mathcal{F}_L$.

Links with optimization algorithms. Our problem shares some common points with the maximization of a deterministic function f (but note that in our case, we only get to see noisy observations of the values of f). When the Lipschitz constant L of f is known, an approximate maximizer can be found with well-known Lipschitz optimization algorithms (e.g., Shubert's algorithm). The case of unknown L has been studied in [JPS93,Hor06]. The DIRECT algorithm of [JPS93] carries out Lipschitz optimization by using the smallest Lipschitz constant that is consistent with the observed data; although it works well in practice, only asymptotic convergence can be guaranteed. The algorithm of [Hor06] iterates over an increasing sequence of possible values of L; under an additional assumption on the minimum increase in the neighborhood of the maximizers, it guarantees a worst-case error of the order of $L^2 T^{-2/d}$ after taking T samples of the deterministic function f.

Adaptation to a global Lipschitz smoothness in bandit problems. We provide in this paper a first step toward a general theory of adaptation. To do so, we focus on the special case of classes $\mathcal{F}_L$ formed by all environments f that are L–Lipschitz with respect to the supremum norm over a subset of $\mathbb{R}^d$: the hypercube $[0,1]^d$ for simplicity. This case covers partially the settings of [BMSS11] and [KSU08], in which the Lipschitz constant was equal to 1, a fact known by the algorithms. (Extensions to Hölderian-type assumptions as in [Kle04,AOS07] will be considered in future work.)

As it is known, getting the minimax-optimal dependency on T is easy, the difficult part is getting that on L without knowing the latter beforehand.

Our contributions. Our algorithm proceeds by discretization as in [Kle04]. To determine the correct discretization step, it first resorts to an uniform exploration yielding a rather crude estimate of the Lipschitz constant (that is however sufficient for our needs); in a second phase, it finds the optimal interval using a standard exploration-exploitation strategy. Our main assumptions are (essentially) that f and its derivative are Lipschitz continuous in the hypercube.

We feel that this two-step approach can potentially be employed in more general settings well beyond ours: with the notation above, the uniform-exploration phase performs a model-selection step and recommends a class $\mathcal{F}_{\widetilde{L}}$, which is used in the second phase to run a continuum-armed bandit strategy tuned with the optimal parameters corresponding to $\widetilde{L} \in \mathcal{L}$. However, for the sake of simplicity, we study only a particular case of this general methodology.

Outline of the paper. In Section 2, we describe the setting and the classes of environments of interest, establish a minimax lower bound on the achievable performance (Section 2.1), and indicate how to achieve it when the global Lips-

chitz parameter L is known (Section 2.2). Our main contribution (Section 3) is then a method to achieve it when the Lipschitz constant is unknown.

2 Setting and notation

We consider a d–dimensional compact set of arms, say, for simplicity, $\mathcal{X} = [0,1]^d$, where $d \geqslant 1$. With each arm $\underline{x} \in [0,1]^d$ is associated a probability distribution $\nu_{\underline{x}}$ with known bounded support, say $[0,1]$; this defines an environment. A key quantity of such an environment is given by the expectations $f(\underline{x})$ of the distributions $\nu_{\underline{x}}$. They define a mapping $f : [0,1]^d \to [0,1]$, which we call the mean-payoff function.

At each round $t \geqslant 1$, the player chooses an arm $\underline{I}_t \in [0,1]^d$ and gets a reward Y_t sampled independently from $\nu_{\underline{I}_t}$ (conditionally on the choice of $\underline{I}_t$). We call a strategy the (possibly randomized) rule that indicates at each round which arm to pull given the history of past rewards.

We write the elements $\underline{x}$ of $[0,1]^d$ in columns; $\underline{x}^{\mathrm{T}}$ will thus denote a row vector with d elements.

Assumption 1 We assume that f is twice differentiable, with Hessians uniformly bounded by M in the following sense: for all $\underline{x} \in [0,1]^d$ and all $\underline{y} \in [0,1]^d$,

$$\left| \underline{y}^{\mathrm{T}} \, H_f(\underline{x}) \, \underline{y} \right| \leqslant M \, \|\underline{y}\|_\infty^2 \, .$$

The ℓ^1–norm of the gradient $\|\nabla f\|_1$ of f is thus continuous and it achieves its maximum on $[0,1]^d$, whose value is denoted by L. As a result, f is Lipschitz with respect to the ℓ^∞–norm with constant L (and L is the smallest[4] constant for which it is Lipschitz): for all $\underline{x}, \underline{y} \in [0,1]^d$,

$$\left| f(\underline{x}) - f(\underline{y}) \right| \leqslant L \, \|\underline{x} - \underline{y}\|_\infty \, .$$

In the sequel we denote by $\mathcal{F}_{L,M}$ the set of environments whose mean-payoff functions satisfy the above assumption. We also denote by $\mathcal{F}_L$ the larger set of environments whose mean-payoff functions f is only constrained to be L–Lipschitz with respect to the ℓ^∞–norm.

2.1 The minimax optimal orders of magnitude of the regret

We denote by

$$f^\star = \sup_{\underline{x} \in [0,1]^d} f(\underline{x}) = \max_{\underline{x} \in [0,1]^d} f(\underline{x})$$

the largest expected payoff in a single round. The expected regret $\overline{R}_T$ at round T is then defined as

$$\overline{R}_T = \mathbb{E}\left[T f^\star - \sum_{t=1}^T Y_t \right] = \mathbb{E}\left[T f^\star - \sum_{t=1}^T f(\underline{I}_t) \right]$$

[4] The proof of the approximation lemma will show why this is the case.

where we used the tower rule and where the expectations are with respect to the random draws of the Y_t according to the $\nu_{\underline{I}_t}$ as well as to any auxiliary randomization the strategy uses.

In this article, we are interested in controlling the worst-case expected regret over all environments whose mean-payoff functions satisfy Assumption 1, i.e., in controlling $\sup_{\mathcal{F}_{L,M}} \overline{R}_T$. The following minimax lower bound follows from a straightforward adaptation of the proof of [BMSS11, Theorem 13], which is provided in Section A in appendix. (The adaptation is needed because the hypothesis on the packing number is not exactly satisfied in the form stated therein.)

Theorem 1. *For all strategies of the player and for all*

$$T \geqslant \max\left\{ L^d,\ \left(\frac{0.15\, L^{2/(d+2)}}{\max\{d,2\}} \right)^d \right\},$$

the worst-case regret over the set $\mathcal{F}_{L,M}$ of all environments that are L–Lipschitz with respect to the ℓ^∞–norm is larger than

$$\sup_{\mathcal{F}_L} \overline{R}_T \geqslant 0.15\, L^{d/(d+2)}\, T^{(d+1)/(d+2)}.$$

The multiplicative constants are not optimized in this bound (a more careful proof might lead to a larger constant in the lower bound).

2.2 How to achieve a minimax optimal regret when L is known

Our aim is to design strategies with worst-case expected regret $\sup_{\mathcal{F}_{L,M}} \overline{R}_T$ less than something of order $L^{d/(d+2)}\, T^{(d+1)/(d+2)}$ when L is unknown. A simple way to do so when L is known was essentially proposed in the introduction of [Kle04] (in the case $d = 1$); it proceeds by discretizing the arm space. The argument is reproduced below and can be used even when L is unknown to recover the optimal dependency $T^{(d+1)/(d+2)}$ on T (but then, with a suboptimal dependency on L).

We consider the approximations $\overline{f}_m$ of f with m^d regular hypercube bins in the ℓ^∞–norm, i.e., m bins are formed in each direction and combined to form the hypercubes. Each of these hypercube bins is indexed by an element $\underline{k} = (k_1, \ldots, k_d) \in \{0, \ldots, m-1\}^d$. The average value of f over the bin indexed by $\underline{k}$ is denoted by

$$\overline{f}_m(\underline{k}) = m^d \int_{\underline{k}/m + [0,1/m]^d} f(\underline{x})\, \mathrm{d}\underline{x}.$$

We then consider the following two-stage strategy, which is based on some strategy MAB for multi-armed bandits; MAB will refer to a generic strategy but we will instantiate below the obtained bound. Knowing L and assuming that T is fixed and known in advance, we may choose beforehand $m = \lceil L^{2/(d+2)} T^{1/(d+2)} \rceil$.

The decomposition of $[0,1]^d$ into m^d bins thus obtained will play the role of the finitely many arms of the multi-armed bandit problem. At round $t \geqslant 1$, whenever the MAB strategy prescribes to pull bin $\underline{K}_t \in \{0,\dots,m-1\}^d$, then first, an arm $\underline{I}_t$ is pulled at random in the hypercube $\underline{K}_t/m + [0,1/m]^d$; and second, given $\underline{I}_t$, the reward Y_t is drawn at random according to $\nu_{\underline{I}_t}$. Therefore, given $\underline{K}_t$, the reward Y_t has an expected value of $\overline{f}_m(\underline{K}_t)$. Finally, the reward Y_t is returned to the underlying MAB strategy.

Strategy MAB is designed to control the regret with respect to the best of the m^d bins, which entails that

$$\mathbb{E}\left[T \max_{\underline{k}} \overline{f}_m(\underline{k}) - \sum_{t=1}^{T} Y_t\right] \leqslant \psi\big(T, m^d\big)\,,$$

for some function ψ that depends on MAB. Now, whenever f is L–Lipschitz with respect to the ℓ^∞–norm, we have that for all $\underline{k} \in \{0,\dots,m-1\}^d$ and all $\underline{x} \in \underline{k}/m + [0,1/m]^d$, the difference $\big|f(\underline{x}) - \overline{f}_m(\underline{k})\big|$ is less than L/m; so that

$$\max_{\underline{x}\in[0,1]^d} f(x) - \max_{\underline{k}} \overline{f}_m(\underline{k}) \leqslant \frac{L}{m}\,.$$

All in all, for this MAB-based strategy, the regret is bounded by the sum of the approximation term L/m and of the regret term for multi-armed bandits,

$$\begin{aligned}\sup_{\mathcal{F}_L} \overline{R}_T &\leqslant T\frac{L}{m} + \psi(T, m^d) \\ &\leqslant L^{d/(d+2)} T^{(d+1)/(d+2)} + \psi\Big(T, \big(\lceil L^{2/(d+2)} T^{1/(d+2)} \rceil\big)^d\Big)\,. \quad (1)\end{aligned}$$

We now instantiate this bound.

The INF strategy of [AB10] (see also [ABL11]) achieves $\psi(T, m') = 2\sqrt{2Tm'}$ and this entails a final $O\big(L^{d/(d+2)}\, T^{(d+1)/(d+2)}\big)$ bound in (1). Note that for the EXP3 strategy of [ACBFS02] or the UCB strategy of [ACBF02], extra logarithmic terms of the order of $\ln T$ would appear in the bound.

3 Achieving a minimax optimal regret not knowing L

In this section, our aim is to obtain a worst-case regret of the minimax-optimal order of $L^{d/(d+2)}\, T^{(d+1)/(d+2)}$ even when L is unknown. To do so, it will be useful to first estimate L; we will provide a (rather crude) estimate suited to our needs, as our goal is the minimization of the regret rather than the best possible estimation of L. Our method is based on the following approximation results.

For the estimation to be efficient, it will be convenient to restrict our attention to the subset $\mathcal{F}_{L,M}$ of $\mathcal{F}_L$, i.e., we will consider the additional assumptions on the existence and boundedness of the Hessians asserted in Assumption 1.

3.1 Some preliminary approximation results

We still consider the approximations $\overline{f}_m$ of f over $[0,1]^d$ with m^d regular bins. We then introduce the following approximation of L:

$$\overline{L}_m = m \max_{\underline{k}\in\{1,\dots,m-2\}^d} \max_{\underline{s}\in\{-1,1\}^d} \left|\overline{f}_m(\underline{k}) - \overline{f}_m(\underline{k}+\underline{s})\right|.$$

This quantity provides a fairly good approximation of the Lipschitz constant, since $m\left(\overline{f}_m(\underline{k}) - \overline{f}_m(\underline{k}+\underline{s})\right)$ is an estimation of the (average) derivative of f in bin $\underline{k}$ and direction $\underline{s}$.

The lemma below relates precisely $\overline{L}_m$ to L: as m increases, $\overline{L}_m$ converges to L.

Lemma 1. *If $f \in \mathcal{F}_{L,M}$ and $m \geqslant 3$, then*

$$L - \frac{7M}{m} \leqslant \overline{L}_m \leqslant L\,.$$

Proof. We note that for all $\underline{k} \in \{1,\dots,m-2\}^d$ and $\underline{s} \in \{-1,1\}^d$, we have by definition

$$\begin{aligned}\left|\overline{f}_m(\underline{k}) - \overline{f}_m(\underline{k}+\underline{s})\right| &= m^d \left|\int_{\underline{k}/m+[0,1/m]^d} \Big(f(\underline{x}) - f\big(\underline{x}+\underline{s}/m\big)\Big)\,\mathrm{d}\underline{x}\right| \\ &\leqslant m^d \int_{\underline{k}/m+[0,1/m]^d} \Big|f(\underline{x}) - f\big(\underline{x}+\underline{s}/m\big)\Big|\,\mathrm{d}\underline{x}\,.\end{aligned}$$

Now, since f is L–Lipschitz in the ℓ^∞–norm, it holds that

$$\Big|f(\underline{x}) - f\big(\underline{x}+\underline{s}/m\big)\Big| \leqslant L\,\big\|\underline{s}/m\big\|_\infty = \frac{L}{m}\,;$$

integrating this bound entails the stated upper bound L on $\overline{L}_m$.

For the lower bound, we first denote by $\underline{x}_\star \in [0,1]^d$ a point such that $\big\|\nabla f(\underline{x}_\star)\big\|_1 = L$. This point belongs to some bin in $\{0,\dots,m-1\}^d$; however, the closest bin $\underline{k}^\star_m$ in $\{1,\dots,m-2\}^d$ is such that

$$\forall \underline{x} \in \underline{k}^\star_m/m + [0,1/m]^d, \qquad \|\underline{x} - \underline{x}_\star\|_\infty \leqslant \frac{2}{m}. \tag{2}$$

Note that this bin $\underline{k}^\star_m$ is such that all $\underline{k}^\star_m + \underline{s}$ belong to $\{0,\dots,m-1\}^d$ and hence legally index hypercube bins, when $\underline{s} \in \{-1,1\}^d$. Now, let $\underline{s}^\star_m \in \{-1,1\}^d$ be such that

$$\nabla f(\underline{x}_\star)\cdot \underline{s}^\star_m = \big\|\nabla f(\underline{x}_\star)\big\|_1 = L\,, \tag{3}$$

where $\cdot$ denotes the inner product in $\mathbb{R}^d$. By the definition of $\overline{L}_m$ as some maximum,

$$\begin{aligned}\overline{L}_m &\geqslant m\left|\overline{f}_m\big(\underline{k}^\star_m\big) - \overline{f}_m\big(\underline{k}^\star_m + \underline{s}^\star_m\big)\right| \\ &= m \times m^d \left|\int_{\underline{k}^\star_m/m+[0,1/m]^d} \Big(f(\underline{x}) - f\big(\underline{x}+\underline{s}^\star_m/m\big)\Big)\,\mathrm{d}\underline{x}\right|. \end{aligned} \tag{4}$$

Now, Taylor's theorem (in the mean-value form for real-valued twice differentiable functions of possibly several variables) shows that for any $\underline{x} \in \underline{k}^\star_m/m + [0,1/m]^d$, there exists two elements ξ and ζ, belonging respectively to the segments between $\underline{x}$ and $\underline{x}_\star$, on the one hand, between $\underline{x}_\star$ and $\underline{x} + \underline{s}^\star_m/m$ on the other hand, such that

$$\begin{aligned}
&f(\underline{x}) - f\big(\underline{x} + \underline{s}^\star_m/m\big) \\
&= \big(f(\underline{x}) - f(\underline{x}_\star)\big) + \Big(f(\underline{x}_\star) - f\big(\underline{x} + \underline{s}^\star_m/m\big)\Big) \\
&= \nabla f(\underline{x}_\star) \cdot (\underline{x} - \underline{x}_\star) + \frac{1}{2}(\underline{x} - \underline{x}_\star)^{\mathrm{T}} \, H_f(\xi) \, (\underline{x} - \underline{x}_\star) \\
&\quad - \nabla f(\underline{x}_\star) \cdot \big(\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big) - \frac{1}{2}\big(\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big)^{\mathrm{T}} \, H_f(\zeta) \, \big(\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big) \\
&= -\nabla f(\underline{x}_\star) \cdot \frac{\underline{s}^\star_m}{m} + \frac{1}{2}(\underline{x} - \underline{x}_\star)^{\mathrm{T}} \, H_f(\xi) \, (\underline{x} - \underline{x}_\star) \\
&\qquad\qquad - \frac{1}{2}\big(\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big)^{\mathrm{T}} \, H_f(\zeta) \, \big(\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big) \, .
\end{aligned}$$

Using (3) and substituting the bound on the Hessians stated in Assumption 1, we get

$$f(\underline{x}) - f\big(\underline{x} + \underline{s}^\star_m/m\big) \leqslant -L + \frac{M}{2}\|\underline{x} - \underline{x}_\star\|_\infty^2 + \frac{M}{2}\big\|\underline{x} + \underline{s}^\star_m/m - \underline{x}_\star\big\|_\infty^2 ;$$

substituting (2), we get

$$f(\underline{x}) - f\big(\underline{x} + \underline{s}^\star_m/m\big) \leqslant -L + \frac{M}{2m^2}\big(2^2 + 3^2\big) \leqslant -L + \frac{7M}{m} \leqslant 0 \, ,$$

where the last inequality holds with no loss of generality (if it does not, then the lower bound on $\overline{L}_m$ in the statement of the lemma is trivial). Substituting and integrating this equality in (4) and using the triangle inequality, we get

$$\overline{L}_m \geqslant L - \frac{7M}{m}.$$

This concludes the proof. □

3.2 A strategy in two phases

Our strategy is described in Figure 1; several notation that will be used in the statements and proofs of some results below are defined therein. Note that we proceed into two phases: a pure exploration phase, when we estimate L, and an exploration–exploitation phase, when we use a strategy designed for the case of finitely-armed bandits on a discretized version of the arm space.

The first step in the analysis is to relate $\widehat{L}_m$ to the quantity it is estimating, namely $\overline{L}_m$.

Parameters:

- Number T of rounds;
- Number m of bins (in each direction) considered in the pure exploration phase;
- Number E of times each of them must be pulled;
- A multi-armed bandit strategy MAB (taking as inputs a number m^d of arms and possibly other parameters).

Pure exploration phase:

1. For each $\underline{k} \in \{0, \dots, m-1\}^d$
   - pull E arms independently uniformly at random in $\underline{k}/m + [0, 1/m]^d$ and get E associated rewards $Z_{\underline{k},j}$, where $j \in \{1, \dots, E\}$;
   - compute the average reward for bin $\underline{k}$,

$$\widehat{\mu}_{\underline{k}} = \frac{1}{E} \sum_{j=1}^{E} Z_{\underline{k},j} \, ;$$

2. Set

$$\widehat{L}_m = m \max_{\underline{k} \in \{1,\dots,m-2\}^d} \max_{\underline{s} \in \{-1,1\}^d} \left| \widehat{\mu}_{\underline{k}} - \widehat{\mu}_{\underline{k}+\underline{s}} \right|$$

   and define $\widetilde{L}_m = \widehat{L}_m + m\sqrt{\frac{2}{E} \ln(2m^d T)}$ as well as $\widetilde{m} = \left\lceil \widetilde{L}_m^{2/(d+2)} T^{1/(d+2)} \right\rceil$.

Exploration–exploitation phase:
Run the strategy MAB with $\widetilde{m}^d$ arms as follows; for all $t = Em + 1, \dots, T$,

1. If MAB prescribes to play arm $\underline{K}_t \in \left\{0, \dots, \widetilde{m} - 1\right\}^d$, pull an arm $\underline{I}_t$ at random in $\underline{K}_t/m + [0, 1/m]^d$;
2. Observe the associated payoff Y_t, drawn independently according to $\nu_{\underline{I}_t}$;
3. Return Y_t to the strategy MAB.

Fig. 1. The considered strategy.

Lemma 2. *With probability at least* $1-\delta$*,*

$$\left|\widehat{L}_m - \overline{L}_m\right| \leqslant m\sqrt{\frac{2}{E}\ln\frac{2m^d}{\delta}}\,.$$

Proof. We consider first a fixed $\underline{k} \in \{0,\dots,m-1\}^d$; as already used in Section 2.2, the $Z_{\underline{k},j}$ are independent and identically distributed according to a distribution on $[0,1]$ with expectation $\overline{f}_m(\underline{k})$, as j varies between 1 and E. Therefore, by Hoeffding's inequality, with probability at least $1-\delta/m^d$

$$\left|\widehat{\mu}_{\underline{k}} - \overline{f}_m(\underline{k})\right| \leqslant \sqrt{\frac{1}{2E}\ln\frac{2m^d}{\delta}}\,.$$

Performing a union bound and using the triangle inequality, we get that with probability at least $1-\delta$,

$$\forall \underline{k},\underline{k}' \in \{0,\dots,m-1\}, \qquad \Big|\,\left|\widehat{\mu}_{\underline{k}} - \widehat{\mu}_{\underline{k}'}\right| - \left|\overline{f}_m(\underline{k}) - \overline{f}_m(\underline{k}')\right|\,\Big| \leqslant \sqrt{\frac{2}{E}\ln\frac{2m^d}{\delta}}\,.$$

This entails the claimed bound. □

By combining Lemmas 1 and 2, we get the following inequalities on $\widetilde{L}_m$, since the latter is obtained from $\widehat{L}_m$ by adding a deviation term.

Corollary 1. *If* $f \in \mathcal{F}_{L,M}$ *and* $m \geqslant 3$*, then, with probability at least* $1-1/T$*,*

$$L - \frac{7M}{m} \leqslant \widetilde{L}_m \leqslant L + 2m\sqrt{\frac{2}{E}\ln\bigl(2m^dT\bigr)}\,.$$

We state a last intermediate result; it relates the regret of the strategy of Figure 1 to the regret of the strategy MAB that it takes as a parameter.

Lemma 3. *Let* $\psi(T',m')$ *be a distribution-free upper bound on the expected regret of the strategy MAB, when run for* T' *rounds on a multi-armed bandit problem with* m' *arms, to which payoff distributions over* $[0,1]$ *are associated. The expected regret of the strategy defined in Figure 1 is then bounded from above as*

$$\sup_{\mathcal{F}_{L,M}} \overline{R}_T \leqslant Em^d + \mathbb{E}\left[\frac{LT}{\widetilde{m}} + \psi\bigl(T - Em^d,\, \widetilde{m}^d\bigr)\right].$$

Proof. As all payoffs lie in $[0,1]$, the regret during the pure exploration phase is bounded by the total length Em^d of this phase.

Now, we bound the (conditionally) expected regret of the MAB strategy during the exploration–exploitation phase; the conditional expectation is with respect to the pure exploration phase and is used to fix the value of $\widetilde{m}$. Using the same arguments as in Section 2.2, the regret during this phase, which lasts $T - Em^d$ rounds, is bounded against any environment in $\mathcal{F}_{L,M}$ by

$$L\frac{T - Em^d}{m} + \psi\bigl(T - Em^d,\, \widetilde{m}^d\bigr)\,.$$

The tower rule concludes the proof. □

We are now ready to state our main result.

Theorem 2. *When used with the multi-armed strategy INF, the strategy of Figure 1 ensures that*

$$\sup_{\mathcal{F}_{L,M}} \overline{R}_T \leqslant T^{(d+1)/(d+2)} \left(9\, L^{d/(d+2)} + 5 \left(2m \sqrt{\frac{2}{E} \ln\bigl(2T^{d+1}\bigr)} \right)^{d/(d+2)} \right) + E m^d + 2\sqrt{2T d^d} + 1 \quad (5)$$

as soon as

$$m \geqslant \frac{8M}{L}\,. \quad (6)$$

In particular, for

$$0 < \gamma < \frac{d(d+1)}{(3d+2)(d+2)} \qquad \textit{and} \qquad \alpha = \frac{1}{d+2} \left(\frac{d+1}{d+2} - \gamma\, \frac{3d+2}{d} \right) > 0\,, \quad (7)$$

the choices of

$$m = \lfloor T^\alpha \rfloor \qquad \textit{and} \qquad E = m^2 \left\lceil T^{2\gamma(d+2)/d} \right\rceil,$$

yield the bound

$$\sup_{\mathcal{F}_{L,M}} \overline{R}_T \leqslant \max\left\{ \left(\frac{8M}{L} + 1 \right)^{1/\alpha},\ L^{d/(d+2)}\, T^{(d+1)/(d+2)} \bigl(9 + \varepsilon(T,d) \bigr) \right\}, \quad (8)$$

where

$$\varepsilon(T,d) = 5T^{-\gamma} \bigl(\ln(2T^d) \bigr)^{d/(d+2)} + T^{-\gamma} + \frac{2\sqrt{2d^d\, T} + 1}{T^{-(d+1)/(d+2)}}$$

vanishes as T tends to infinity.

Note that the choices of E and m solely depend on T, which may however be unknown in advance; standard arguments, like the doubling trick, can be used to circumvent the issue, at a minor cost given by an additional constant multiplicative factor in the bound.

Remark 1. There is a trade-off between the value of the constant term in the maximum, $(1+8M/L)^{1/\alpha}$, and the convergence rate of the vanishing term $\varepsilon(T,d)$ toward 0, which is of order γ. For instance, in the case $d = 1$, the condition on γ is $0 < \gamma < 2/15$; as an illustration, we get

- a constant term of a reasonable size, since $1/\alpha \leqslant 4.87$, when the convergence rate is small, $\gamma = 0.01$;
- a much larger constant, since $1/\alpha = 60$, when the convergence rate is faster, $\gamma = 2/15 - 0.01$.

Proof. For the strategy INF, as recalled above, $\psi(T', m') = 2\sqrt{2T'm'}$. The bound of Lemma 3 can thus be instantiated as

$$\begin{aligned}
&\sup_{\mathcal{F}_{L,M}} \overline{R}_T \leqslant Em^d + \mathbb{E}\left[\frac{LT}{\widetilde{m}} + 2\sqrt{2T\widetilde{m}^d}\right] \\
&\leqslant Em^d + \mathbb{E}\left[T^{(d+1)/(d+2)}\frac{L}{\widetilde{L}_m^{2/(d+2)}} + 2\sqrt{2T\,e\Big(T^{1/(d+2)}\widetilde{L}_m^{2/(d+2)}\Big)^d} + 2\sqrt{2Td^d}\right].
\end{aligned}$$

To get the last inequality, we substituted the definition

$$\widetilde{m} = \left\lceil \widetilde{L}_m^{2/(d+2)} T^{1/(d+2)} \right\rceil \leqslant \widetilde{L}_m^{2/(d+2)} T^{1/(d+2)} \left(1 + \frac{1}{\widetilde{L}_m^{2/(d+2)} T^{1/(d+2)}}\right)$$

and separated the cases where $\widetilde{L}_m^{2/(d+2)} T^{1/(d+2)}$ is smaller or larger than d. In the first case, we simply bound $\widetilde{m}$ by d and get the $2\sqrt{2Td^d}$ term. When the quantity of interest is larger than d, then we get the central term in the expectation above by using the fact that $(1+1/x)^d \leqslant (1+1/d)^d \leqslant e$ whenever $x \geqslant d$.

We will now use the lower and upper bounds on $\widehat{L}_m$ stated by Corollary 1. In the sequel we will make repeated uses of the following convexity inequality: for all integers p, all $u_1, \ldots, u_p > 0$, and all $\alpha \in [0,1]$,

$$\big(u_1 + \ldots + u_p\big)^\alpha \leqslant p^{\alpha-1}\big(u_1^\alpha + \ldots + u_p^\alpha\big) \leqslant u_1^\alpha + \ldots + u_p^\alpha\,. \tag{9}$$

By resorting to (9), we get that with probability at least $1 - 1/T$,

$$\begin{aligned}
&2\sqrt{2T\,e\Big(T^{1/(d+2)}\widetilde{L}_m^{2/(d+2)}\Big)^d} \\
&= 2\sqrt{2e}\; T^{(d+1)/(d+2)}\; \widetilde{L}_m^{d/(d+2)} \\
&\leqslant 2\sqrt{2e}\; T^{(d+1)/(d+2)} \left(L + 2m\sqrt{\frac{2}{E}\ln\big(2m^dT\big)}\right)^{d/(d+2)} \\
&\leqslant 2\sqrt{2e}\; T^{(d+1)/(d+2)} \left(L^{d/(d+2)} + \left(2m\sqrt{\frac{2}{E}\ln\big(2m^dT\big)}\right)^{d/(d+2)}\right).
\end{aligned}$$

On the other hand, with probability at least $1 - 1/T$,

$$\widetilde{L}_m \geqslant L - \frac{7M}{m} \geqslant \frac{L}{8}\,,$$

where we assumed that m and E are chosen large enough for the lower bound of Corollary 1 to be larger than $L/8$. This is indeed the case as soon as

$$\frac{7M}{m} \leqslant \frac{7L}{8}\,, \qquad \text{that is,} \qquad m \geqslant \frac{8M}{L}\,,$$

which is exactly the condition (6).

Putting all things together (and bounding m by T in the logarithm), with probability at least $1 - 1/T$, the regret is less than

$$Em^d \;+\; T^{(d+1)/(d+2)} \frac{L}{(L/8)^{2/(d+2)}} \;+\; 2\sqrt{2Td^d}$$
$$+\, 2\sqrt{2e}\; T^{(d+1)/(d+2)} \left(L^{d/(d+2)} + \left(2m \sqrt{\frac{2}{E} \ln\bigl(2T^{d+1}\bigr)} \right)^{d/(d+2)} \right) ; \quad (10)$$

on the event of probability smaller than $\delta = 1/T$ where the above bound does not necessarily hold, we upper bound the regret by T. Therefore, the expected regret is bounded by (10) plus 1. Bounding the constants as

$$8^{2/(d+2)} = 4 \qquad \text{and} \qquad 2\sqrt{2e} \leqslant 5$$

concludes the proof of the first part of the theorem.

The second part follows by substituting the values of E and m in the expression above and by bounding the regret by T_0 for the time steps $t \leqslant T_0$ for which the condition (6) is not satisfied.

More precisely, the bound obtained in the first part shows that it is necessary that $E \gg m^2$ and $Em^d \ll T^{(d+1)/(d+2)}$ for the goal mentioned in the introduction to be achieved. This is why we looked for suitable values of m and E in the following form:

$$m = \lfloor T^\alpha \rfloor \qquad \text{and} \qquad E = m^2 \left\lceil T^{2\gamma(d+2)/d} \right\rceil ,$$

where α and γ are positive. We choose α as a function of γ so that the terms

$$Em^d = \bigl(\lfloor T^\alpha \rfloor \bigr)^{d+2} \left\lceil T^{2\gamma(d+2)/d} \right\rceil$$
$$\text{and} \qquad T^{(d+1)/(d+2)} \left(\frac{m}{\sqrt{E}} \right)^{d/(d+2)} = T^{(d+1)/(d+2)} \Bigl(\left\lceil T^{2\gamma(d+2)/d} \right\rceil \Bigr)^{-d/(2(d+2))}$$

are approximatively balanced; for instance, such that

$$\alpha(d+2) + 2\gamma(d+2)/d = (d+1)/(d+2) - \gamma ,$$

which yields the proposed expression (7). The fact that α needs to be positive entails the constraint on γ given in (7).

When condition (6) is met, we substitute the values of m and E into (5) to obtain the bound (8); the only moment in this substitution when taking the upper or lower integer parts does not help is for the term Em^d, for which we write (using that $T \geqslant 1$)

$$Em^d = m^{d+2} \left\lceil T^{2\gamma(d+2)/d} \right\rceil \leqslant T^{\alpha(d+2)} \bigl(1 + T^{2\gamma(d+2)/d} \bigr)$$
$$\leqslant 2\, T^{\alpha(d+2)} T^{2\gamma(d+2)/d} = 2\, T^{(d+1)/(d+2) - \gamma} .$$

When condition (6) is not met, which can only be the case when T is such that $T^\alpha < 1 + 8M/L$, that is, $T < T_0 = (1 + 8M/L)^{1/\alpha}$, we upper bound the regret by T_0. □

Acknowledgements

This work was supported in part by French National Research Agency (ANR, project EXPLO-RA, ANR-08-COSI-004) and the PASCAL2 Network of Excellence under EC grant no. 216886.

References

AB10. J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

ABL11. J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory*. Omnipress, 2011.

ACBF02. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002.

ACBFS02. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Agr95. R. Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33:1926–1951, 1995.

AOS07. P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 454–468, 2007.

BM10. S. Bubeck and R. Munos. Open-loop optimistic planning. In *Proceedings of the 23rd Annual Conference on Learning Theory*. Omnipress, 2010.

BMSS11. S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. $\mathcal{X}$–armed bandits. *Journal of Machine Learning Research*, 2011. To appear.

Cop09. E. Cope. Regret and convergence bounds for immediate-reward reinforcement learning with continuous action spaces. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

DHK08. V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366. Omnipress, 2008.

Hor06. Matthias Horn. Optimal algorithms for global optimization in case of unknown Lipschitz constant. *Journal of Complexity*, 22(1), 2006.

JPS93. D.R. Jones, C.D. Perttunen, and B.E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

Kle04. R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, 2004.

KSU08. 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.

Rob52. H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

WAM09. Y. Wang, J.Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2009.

YM11. J.Y. Yu and S. Mannor. Unimodal bandits. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

A Proof of Theorem 1

Proof. We slightly adapt the (end of the) proof of [BMSS11, Theorem 13]; we take the metric $\ell(\underline{x}, \underline{y}) = L\|\underline{x} - \underline{y}\|_\infty$. For $\varepsilon \in (0, 1/2)$, the ε–packing number of $[0,1]^d$ with respect to ℓ equals

$$\mathcal{N}\big([0,1]^d,\, \ell,\, \varepsilon\big) = \big(\lfloor L/\varepsilon \rfloor\big)^d \geqslant 2$$

provided that $L/\varepsilon \geqslant 2$, that is, $\varepsilon \leqslant L/2$. Therefore, Step 5 of the mentioned proof shows that

$$\sup_{\mathcal{F}_{L,M}} \overline{R}_T \geqslant T\varepsilon \left(0.5 - 2.2\,\varepsilon \sqrt{\frac{T}{\big(\lfloor L/\varepsilon \rfloor\big)^d}}\right)$$

for all $0 < \varepsilon < \min\{1, L\}/2$. We now optimize this bound.

Whenever $L/\varepsilon \geqslant \max\{d, 2\}$, we have

$$\lfloor L/\varepsilon \rfloor \geqslant L/\varepsilon - 1 \geqslant \frac{1}{2}\, L/\varepsilon$$

in the case where $d = 1$, while for $d \geqslant 2$,

$$\big(\lfloor L/\varepsilon \rfloor\big)^d \geqslant \big(L/\varepsilon - 1\big)^d \geqslant \frac{1}{4}\big(L/\varepsilon\big)^d,$$

where we used the fact that $(1 - 1/x)^d \geqslant (1 - 1/d)^d \geqslant (1 - 1/2)^2 = 1/4$ for all $x \geqslant d$ and $d \geqslant 2$.

Therefore, whenever $0 < \varepsilon < \min\big\{1/2,\, L/d,\, L/2\big\}$,

$$\sup_{\mathcal{F}_{L,M}} \overline{R}_T \geqslant T\varepsilon \left(0.5 - 4.4\,\varepsilon^{1+d/2} \sqrt{\frac{T}{L^d}}\right).$$

We take ε of the form

$$\varepsilon = \gamma\, L^{d/(d+2)}\, T^{-1/(d+2)}$$

for some constant $\gamma < 1$ to be defined later on. The lower bound then equals

$$\gamma\, L^{d/(d+2)}\, T^{(d+1)/(d+2)}\big(0.5 - 4.4\,\gamma^{1+d/2}\big) \geqslant \gamma\, L^{d/(d+2)}\, T^{(d+1)/(d+2)}\big(0.5 - 4.4\,\gamma^{3/2}\big)$$

where we used the fact that $\gamma < 1$ and $d \geqslant 1$ for the last inequality. Taking γ such that $0.5 - 4.4\,\gamma^{3/2} = 1/4$, that is, $\gamma = 1/(4 \times 4.4)^{2/3} \geqslant 0.14$, we get the stated bound.

It only remains to see that the indicated condition on T proceeds from the value of ε provided above, the constraint $\varepsilon < \min\big\{1/2,\, L/d,\, L/2\big\}$, and the upper bound $\gamma \geqslant 0.15$. □