Online Optimization in $\mathcal{X}$-Armed Bandits

Sébastien Bubeck
INRIA Lille, SequeL project, France
sebastien.bubeck@inria.fr

Rémi Munos
INRIA Lille, SequeL project, France
remi.munos@inria.fr

Gilles Stoltz
Ecole Normale Supérieure and HEC Paris
gilles.stoltz@ens.fr

Csaba Szepesvári
Department of Computing Science, University of Alberta
szepesva@cs.ualberta.ca

1 Introduction and motivation

Bandit problems with continuous set of arms arise in many settings. Of particular interest for this workshop are the cases of on-line parameter tuning or optimization of controllers based on simulations.

We formalize the problem as follows. We consider an arbitrary set $\mathcal{X}$ (we do not talk about measurability issues in this extended abstract), 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 random reward (in $[0,1]$) whose distribution depends on the arm chosen. We note $f(x)$ the mean of the reward received while pulling arm x. The goal of the decision maker is to choose the arms so as to maximize the sum of the rewards that he receives. For a sequence $X_1, \ldots, X_n$ of pulled arms we define the cumulative regret as

$$R_n = n \sup_{x \in \mathcal{X}} f(x) - \sum_{t=1}^{n} f(X_t).$$

We can see this problem as an online optimization of the noisy function f on $\mathcal{X}$.

To solve this problem we propose a practical algorithm (anytime and easy to implement) motivated by the recent very successful tree-based optimization algorithms [5; 3; 2] and show that under weak assumptions on f and $\mathcal{X}$ it enjoys a nice growth rate for $\mathbb{E}R_n$. In particular our assumptions on $\mathcal{X}$ are weaker than Auer et al. [1] who considered $\mathcal{X} = \mathbb{R}$ and the assumption on f is weaker than Kleinberg et al. [4] who assumed f to be Lipschitzian on a metric space $\mathcal{X}$.

To be more precise we assume that the decision maker knows a dissimilarity function 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. The assumption on $\mathcal{X}$ is that we are 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.

This allows us to obtain a regret which scales as $\widetilde{O}(\sqrt{n})$ [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

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

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.

Since the algorithm 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 Preliminaries

Dissimilarity: We assume that $\mathcal{X}$ is equipped with a *dissimilarity* ℓ, that is, a non-negative mapping $\ell : \mathcal{X}^2 \to \mathbb{R}$ satisfying $\ell(x, x) = 0$. ℓ will capture the smoothness of the mean-payoff function.

Tree of coverings: We consider a tree of coverings of $\mathcal{X}$ defined as follows :

- (h, i) (for $h \geq 0$ and $1 \leq i \leq 2^h$) is the i–th node of depth h and corresponds to a subset $\mathcal{P}_{h,i} \subset \mathcal{X}$;
- the root corresponds to the whole domain, i.e., $\mathcal{X} = \mathcal{P}_{0,1}$;
- any parent node (h, i) is covered by its two children $(h + 1, 2i - 1)$ and $(h + 1, 2i)$ (note that we will assume that the size of the regions shrinks as the depth increases).
- For any given $h \geq 0$, the $\{\mathcal{P}_{h,i}\}_{1 \leq i \leq 2^h}$ are disjoints.

Example: A typical choice for the coverings in a cubic domain $[0, 1]^d$ 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.

3 The Hierarchical Optimistic Optimization (HOO) algorithm

3.1 Global strategy given B–values for each node

- Start with all nodes of the tree "turned off".
- At each round follow a path from the root to a turned-off node (h, i), where at each node along the path we select the child (ties broken arbitrarily) with larger B–value (the B–values are defined below).
- Pull a point in $\mathcal{P}_{h,i}$ and turn on (h, i).

3.2 Definition of B–values

- Let $N_{h,i}(n)$ be the number of times we followed a path going through (h, i).
- Let $\widehat{\mu}_{h,i}(n)$ be the empirical average of rewards collected when we followed a path going through (h, i).
- We consider the following upper confidence bound for each turned on node :

$$U_{h,i}(n) = \widehat{\mu}_{h,i}(n) + \sqrt{\frac{2 \ln n}{N_{h,i}(n)}} + \text{ diam}(\mathcal{P}_{h,i}),$$

  where $\text{diam}(\mathcal{P}_{h,i}) = \sup_{x,y \in \mathcal{P}_{h,i}} l(x, y)$.
- Then for turned-off nodes we set the B–values to be infinite and for a turned-on node :

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

4 Assumptions and statement of the main result

The first assumption concerns the function we want to optimize.

Assumption 1. *The mean-payoff f is* weakly Lipschitz *with respect to ℓ, i.e. for all $x, y \in \mathcal{X}$,*

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

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.

The second assumption is closely related to Assumption 2 of Auer et al. [1], 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 $\varepsilon \to 0$. Here, we capture this by defining a new notion, the near-optimality dimension of the mean-payoff function (which is also related to the zooming dimension defined by Kleinberg et al. [4]).

Assumption 2. *Let $\mathcal{X}_\varepsilon = \{x \in \mathcal{X}, f^* - f(x) \leq \varepsilon\}$ be the set of ε–optimal arms. We assume that $\mathcal{X}_\varepsilon$ can be packed with $O(\varepsilon^{-d})$ balls of radius ε. We say that d is the near-optimality dimension of the mean-payoff function.*

Under these assumptions and if $\mathrm{diam}(\mathcal{P}_{h,i})$ tends to 0 at geometric rate with h, we prove the following result:[2]

Theorem 1 (Main result)**.** *There exists a constant $C(d)$ such that for all $n \geq 1$,*

$$\mathbb{E}R_n \leq C(d)\, n^{(d+1)/(d+2)} \left(\ln n\right)^{1/(d+2)}\,.$$

Remark 1. *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 $x, y \in \mathcal{X}_\varepsilon$.*

5 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):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*. \tag{2}$$

Under this assumption, the result of Auer et al. [1] shows that for $D = 1$, the regret is $\Theta(\sqrt{n})$. 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 1) and that the near-optimality dimension is 0. Thus our regret is $\widetilde{O}(\sqrt{n})$, i.e., the rate is independent of the dimension D.

[2] whose precise statement as well as its proof are provided in the full-length paper but not referenced here to preserve anonymity

In comparison, Kleinberg et al. [4] have to satisfy a global Lipschitz assumption, thus 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 regret becomes $\widetilde{O}(n^{(D(\alpha-1)+\alpha)/(D(\alpha-1)+2\alpha)})$, which is strictly worse than $\widetilde{O}(\sqrt{n})$ and in fact becomes close to the slow rate $\widetilde{O}(n^{(D+1)/(D+2)})$ 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 $\widetilde{O}(\sqrt{n})$, 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.

References

[1] 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.

[2] P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Proceedings of 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[3] S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in Monte-Carlo go. Technical Report RR-6062, INRIA, 2006.

[4] 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.

[5] L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.