

Lipschitz Bandits without the Lipschitz Constant

Sébastien Bubeck¹, Gilles Stoltz², Jia Yuan Yu³

¹ Princeton University

² CNRS — École normale supérieure — INRIA & HEC Paris

³ IBM Research, Dublin



Continuum-armed bandit problems

Framework and (partial) state of the art

We consider an arbitrary set \mathcal{X} .

With each arm $\underline{x} \in \mathcal{X}$ is associated an (unknown) probability distribution $\nu_{\underline{x}}$ with known bounded support, say $[0, 1]$.

This defines the (unknown) **environment**.

At each round $t \geq 1$,

- the gambler chooses an arm $\underline{l}_t \in \mathcal{X}$,
- he then gets a reward Y_t sampled independently from $\nu_{\underline{l}_t}$ (conditionally on the choice of \underline{l}_t).

A **strategy** is a way to pick (possibly at random) \underline{l}_t solely based on the rewards Y_1, \dots, Y_{t-1} obtained in the past.

Now, what is a **good** strategy?

Summary of what we have seen so far

- The gambler repeatedly picks arms $I_t \in \mathcal{X}$ based on the past rewards Y_1, \dots, Y_{t-1} and gets rewards Y_t sampled from ν_{I_t} .
-

The aim of the gambler is to obtain a cumulative payoff $Y_1 + \dots + Y_T$ as large as possible.

We denote by $f(\underline{x})$ the expectation of the distribution $\nu_{\underline{x}}$, for $\underline{x} \in \mathcal{X}$.

We call $f : \mathcal{X} \rightarrow [0, 1]$ the **mean-payoff function**.

Most results in the literature are about the **expected** cumulative payoff,

$$\mathbb{E} \left[\sum_{t=1}^T Y_t \right] = \mathbb{E} \left[\sum_{t=1}^T f(I_t) \right]$$

Summary of what we have seen so far

- The gambler repeatedly picks arms $I_t \in \mathcal{X}$ based on the past rewards Y_1, \dots, Y_{t-1} and gets rewards Y_t sampled from ν_{I_t} .
 - We denote by $f(\underline{x})$ the expectation of the distribution $\nu_{\underline{x}}$.
-

The aim of the gambler is to maximize his expected cumulative payoff,

$$\mathbb{E} \left[\sum_{t=1}^T Y_t \right] = \mathbb{E} \left[\sum_{t=1}^T f(I_t) \right]$$

Or, equivalently, to minimize his (expected cumulative) **regret**

$$R_T = \mathbb{E} \left[\sum_{t=1}^T (f^* - f(I_t)) \right],$$

where $f^* = \sup_{\underline{x} \in \mathcal{X}} f(\underline{x})$

Summary of what we have seen so far

- The gambler repeatedly picks arms $I_t \in \mathcal{X}$ based on the past rewards Y_1, \dots, Y_{t-1} and gets rewards Y_t sampled from ν_{I_t} .
 - We denote by $f(\underline{x})$ the expectation of the distribution $\nu_{\underline{x}}$
 - and by $f^* = \sup_{\underline{x} \in \mathcal{X}} f(\underline{x})$ the expected payoff of the best arm.
-

We consider first the case when $\mathcal{X} = \{1, \dots, K\}$ is **finite**.

The regret of the **INF** strategy of Audibert and Bubeck '10 is **uniformly** controlled over all environments as

$$\sup_{\nu_1, \dots, \nu_K} R_T = \sup_{\nu_1, \dots, \nu_K} \mathbb{E} \left[\sum_{t=1}^T (f^* - f(I_t)) \right] \leq 2\sqrt{2} \sqrt{TK}$$

The orders of magnitude in T and K are **minimax optimal**, as follows from the lower bound exhibited in Auer et al. '02:

No strategy can **uniformly** achieve better orders.

When \mathcal{X} is larger, no uniform control of the regret with respect to **all possible environments** can be achieved in general.

For the sake of concreteness, we consider $\mathcal{X} = [0, 1]^d$ in the sequel.

Assumptions are needed on the environments and they are often stated via **regularity** assumptions on the **mean-payoff** functions f .

The uniform regret bounds exhibited in the literature are of the form: for algorithms with (possibly vector-valued) parameters p ,

$$\sup_{f \in \mathcal{F}(p)} R_T(f) \leq \gamma_{T,p} \quad \text{where} \quad \gamma_{T,p} = o(T)$$

For instance, the class of environments for the **CAB1** algorithm of Kleinberg '04 is formed by environments that are (α, L, δ) -uniformly locally Lipschitz.

The strategy CAB1 needs to **know α** to get the optimal dependency in T of the regret bound.

When \mathcal{X} is larger, no uniform control of the regret with respect to **all possible environments** can be achieved in general.

For the sake of concreteness, we consider $\mathcal{X} = [0, 1]^d$ in the sequel.

Assumptions are needed on the environments and they are often stated via **regularity** assumptions on the **mean-payoff** functions f .

The uniform regret bounds exhibited in the literature are of the form: for algorithms with (possibly vector-valued) parameters p ,

$$\sup_{f \in \mathcal{F}(p)} R_T(f) \leq \gamma_{T,p} \quad \text{where} \quad \gamma_{T,p} = o(T)$$

For the **Zooming** algorithm of Kleinberg et al. '08, the comparison class is formed by environments that are **1**-Lipschitz with respect to a fixed and **known metric**.

When \mathcal{X} is larger, no uniform control of the regret with respect to **all possible environments** can be achieved in general.

For the sake of concreteness, we consider $\mathcal{X} = [0, 1]^d$ in the sequel.

Assumptions are needed on the environments and they are often stated via **regularity** assumptions on the **mean-payoff** functions f .

The uniform regret bounds exhibited in the literature are of the form: for algorithms with (possibly vector-valued) parameters p ,

$$\sup_{f \in \mathcal{F}(p)} R_T(f) \leq \gamma_{T,p} \quad \text{where} \quad \gamma_{T,p} = o(T)$$

Similar issues arise for the discretization algorithm studied in Auer et al. '07, the HOO algorithm of Bubeck et al. '11, etc.

When \mathcal{X} is larger, no uniform control of the regret with respect to **all possible environments** can be achieved in general.

For the sake of concreteness, we consider $\mathcal{X} = [0, 1]^d$ in the sequel.

Assumptions are needed on the environments and they are often stated via **regularity** assumptions on the **mean-payoff** functions f .

The uniform regret bounds exhibited in the literature are of the form: for algorithms with (possibly vector-valued) parameters p ,

$$\sup_{f \in \mathcal{F}(p)} R_T(f) \leq \gamma_{T,p} \quad \text{where} \quad \gamma_{T,p} = o(T)$$

The drawback of these bounds is that the class of environments $\mathcal{F}(p)$ which a given algorithm is uniformly competitive with respect to depends on its parameters p .

This is **not the right perspective**: The class of environments comes first and the algorithms should adapt to it!

Adaptation to unknown smoothness parameters

Statement of our goal

(Partial) solution for the adaptation to the Lipschitz constant

By **adaptation** to unknown **smoothness** parameters we mean that based on a family of strategies indexed by the parameters $p \in \mathcal{P}$, with **optimal** regret bounds of the form

$$\sup_{f \in \mathcal{F}(p)} R_T(f) \leq \gamma_{T,p},$$

we should construct a **meta-strategy** such that for all mean-payoff functions f ,

$$R_T(f) \leq \inf_{p: f \in \mathcal{F}(p)} \gamma_{T,p}$$

This meta-strategy should be based on **no parameter**, i.e., should only require the (**minimal!**) prior knowledge of the fact that

$$f \in \bigcup_{p \in \mathcal{P}} \mathcal{F}(p)$$

We thought that the adaptation to an unknown Lipschitz constant L would be a good warm-up.

It turns out that, as confirmed by non-parametric statisticians, this case is rather delicate as the goal is to get a constant multiplicative factor right.

(See the next slide for details.)

The case of the adaptation to the unknown Hölderian order α might have been easier, since there, the goal is to get the order in T right (as a function of α) and the finer study of the multiplicative constants is usually omitted.

But, oh well, let's see what we got for the adaptation to L ...

Definition: We denote by $\mathcal{F}(L)$ the class of the mean-payoff functions $f : [0, 1]^d \rightarrow [0, 1]$ which are L -Lipschitz,

$$\forall \underline{x}, \underline{y} \in [0, 1]^d, \quad |f(\underline{x}) - f(\underline{y})| \leq L \|\underline{x} - \underline{y}\|_\infty$$

Fact 1 (Bubeck et al. '11): For all strategies,

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \geq 0.15 L^{d/(d+2)} T^{(d+1)/(d+2)}$$

Fact 2: Knowing L , decomposing $[0, 1]^d$ into m^d bins and using a two-step algorithm, the regret is less than

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \leq T \frac{L}{m} + 2\sqrt{2 T m^d}$$

(resorting to the INF strategy to choose a bin and then pulling a point at random in the bin).

Definition: We denote by $\mathcal{F}(L)$ the class of the mean-payoff functions $f : [0, 1]^d \rightarrow [0, 1]$ which are L -Lipschitz,

$$\forall \underline{x}, \underline{y} \in [0, 1]^d, \quad |f(\underline{x}) - f(\underline{y})| \leq L \|\underline{x} - \underline{y}\|_\infty$$

Fact 1 (Bubeck et al. '11): For all strategies,

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \geq 0.15 L^{d/(d+2)} T^{(d+1)/(d+2)}$$

Fact 2: Knowing L , decomposing $[0, 1]^d$ into m^d bins and using a two-step algorithm, the regret is less than

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \leq T \frac{L}{m} + 2\sqrt{2 T m^d} = O\left(L^{d/(d+2)} T^{(d+1)/(d+2)}\right)$$

for a choice of m of the order of $L^{2/(d+2)} T^{1/(d+2)}$.

That is, knowing L , the **minimax** orders of magnitude in T and L can be **achieved**.

Definition: We denote by $\mathcal{F}(L)$ the class of the mean-payoff functions $f : [0, 1]^d \rightarrow [0, 1]$ which are L -Lipschitz,

$$\forall \underline{x}, \underline{y} \in [0, 1]^d, \quad |f(\underline{x}) - f(\underline{y})| \leq L \|\underline{x} - \underline{y}\|_\infty$$

Fact 1 (Bubeck et al. '11): For all strategies,

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \geq 0.15 L^{d/(d+2)} T^{(d+1)/(d+2)}$$

Fact 2: Knowing L , decomposing $[0, 1]^d$ into m^d bins and using a two-step algorithm, the regret is less than

$$\sup_{f \in \mathcal{F}(L)} R_T(f) \leq T \frac{L}{m} + 2\sqrt{2 T m^d} = O\left(\max\{L, 1\} T^{(d+1)/(d+2)}\right)$$

not knowing L and taking m of the order of $T^{1/(d+2)}$.

The optimal **multiplicative factor** of $L^{d/(d+2)}$ is replaced by the **larger** quantity $\max\{L, 1\}$.

Our **ideal goal** would thus have been to construct a meta-strategy such that for **all Lipschitz** mean-payoff functions f , with Lipschitz constant denoted by L_f ,

$$R_T(f) \leq \square L_f^{d/(d+2)} T^{(d+1)/(d+2)}$$

where \square is a numerical constant.

The only information to which the meta-strategy has access is that the mean-payoff function f is Lipschitz.

Our meta-strategy

- **estimates** (rather crudely) L_f by \tilde{L}_f in a **pure exploration** phase,
- picks the discretization + INF strategy associated with \tilde{L}_f to perform exploration–exploitation in the **second phase**.

Our **ideal goal** would thus have been to construct a meta-strategy such that for **all Lipschitz** mean-payoff functions f , with Lipschitz constant denoted by L_f ,

$$R_T(f) \leq \square L_f^{d/(d+2)} T^{(d+1)/(d+2)}$$

where \square is a numerical constant.

The only information to which the meta-strategy has access is that the mean-payoff function f is Lipschitz.

This two-stage procedure (1. estimating the parameters crudely; 2. substituting this estimate to pick a base strategy) is probably a **general trick**.

Our **ideal goal** would thus have been to construct a meta-strategy such that for **all Lipschitz** mean-payoff functions f , with Lipschitz constant denoted by L_f ,

$$R_T(f) \leq \square L_f^{d/(d+2)} T^{(d+1)/(d+2)}$$

where \square is a numerical constant.

The only information to which the meta-strategy has access is that the mean-payoff function f is Lipschitz.

However, at this stage I must **confess** that we did not exactly get the bound stated above but a somewhat less sharp one...

Pure exploration phase of length Em^d (parameters m and E)

- For each bin $\underline{k} \in \{0, \dots, m-1\}^d$
 - 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} ,

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

- Set

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

define the upper bound $\tilde{L} = \hat{L} + m \sqrt{\frac{2}{E} \ln(2m^d T)}$

and the associated discretization parameter

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

Exploration–exploitation phase

(with parameter \tilde{m} defined in the previous phase)

Run the strategy INF of Audibert and Bubeck '10 with \tilde{m}^d arms as follows.

For all rounds $t = Em^d + 1, \dots, T$,

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

We could only prove a uniform bound on the restricted class $\mathcal{F}(L, M)$ of mean-payoff functions that are L -Lipschitz and whose Hessians are uniformly bounded by M (in the supremum norm).

Theorem

For well-chosen values of the parameters E and m (solely depending on T), the procedure described in the previous slides satisfies that for all L and M ,

$$\sup_{f \in \mathcal{F}_{L,M}} R_T(f) \leq \max \left\{ C_{M,L}, L^{d/(d+2)} T^{(d+1)/(d+2)} (9 + \varepsilon(T, d)) \right\},$$

where

- $C_{M,L}$ is a constant depending on L and (unfortunately) on M ,
- $\varepsilon(T, d)$ vanishes as $T \rightarrow \infty$.

The **unpleasant** fact in the bound

$$\sup_{f \in \mathcal{F}_{L,M}} R_T(f) \leq \max \left\{ C_{M,L}, L^{d/(d+2)} T^{(d+1)/(d+2)} (9 + \varepsilon(T, d)) \right\}$$

is that $C_{M,L} \rightarrow \infty$ as $M \rightarrow \infty$.

It thus does not yield the **uniform** optimal $\square L^{d/(d+2)} T^{(d+1)/(d+2)}$ control over \mathcal{F}_L .

However...

- It could be worse — the dependency of the bound in M is in the additive form, not in the multiplicative form.
- Is it an artifact of the analysis? — Smarter people in the audience could probably do better than we did.

Conclusion

A good and important problem... which still needs some more work!

We stated a **crucial problem**: getting the algorithms to adapt to the unknown smoothness parameters of the stochastic environments they are facing.

We presented a **partial solution** in the case of the adaptation to the Lipschitz constant — we still needed an unpleasant assumption about Hessians.

We however **suspect** that the two-phase methodology used here could be helpful for other (simpler?) adaptation problems.

E.g., for the adaptation to the **Hölderian order α** , only orders of magnitude in T are of interest (and not so much the multiplicative constants).

In any case: There is **room for improvement** for the motivated ones in the audience!