Polynomial Cost of Adaptation for X-Armed Bandits

Hédi Hadiji, Laboratoire de Mathématiques d'Orsay, Université Paris-Sud

Introduction

X-armed bandits are a generalization of the K-armed bandit problem, in which the action set is continuous. [Locatelli and Carpentier '18] recently uncovered obstacles to designing algorithms that adapt to the regularity of the mean-payoff function.

We revisit the lower bound and provide an algorithm that is as adaptive as possible.

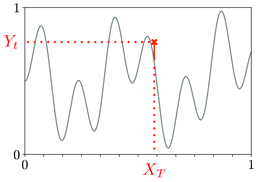
X-Armed Bandits

Arm space $\mathcal{X} = [0,1]$ Unknown mean-payoff function $f \in [0,1]^{\mathcal{X}}$

For $t = 1, \dots, T$:

- pick $X_t \in \mathcal{X}$
- observe and receive reward $Y_t = f(X_t) + \varepsilon_t$

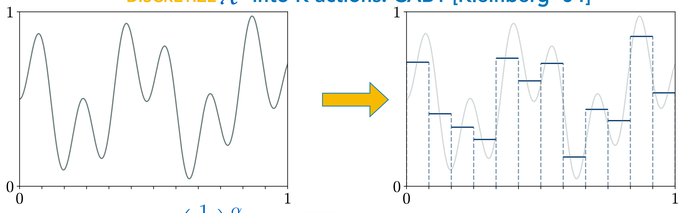
Regret: $R_T = T \max_{x \in \mathcal{X}} f(x) - \mathbb{E}\left[\sum_{t=1}^{T} f(X_t)\right]$



Known smoothness? Discretize

α-Hölder $\quad \forall x, y \in \mathcal{X} \quad |f(x) - f(y)| \leqslant |x - y|^{\alpha}$

DISCRETIZE $\mathcal{X}$ into K actions: CAB1 [Kleinberg '04]



$R_T \leq T\left(\frac{1}{K}\right)^{\alpha} + c\sqrt{KT}$ then tune K according to α

$\min_{\text{alg.}} \max_{f\ \alpha\text{-Hölder}} R_T \asymp T^{(\alpha+1)/(2\alpha+1)}$

Unknown smoothness?

Dream is *full adaptation* at no cost: getting the same guarantees as if smoothness was known

Model selection? Cross-validation? Exploration is costly!

Theorem : No full adaptation [Locatelli and Carpentier '18]

If $\alpha \leqslant \gamma$ and $\max_{f\ \gamma\text{-Hölder}} R_T \leqslant B$ then $\max_{f\ \alpha\text{-Hölder}} R_T \geqslant cTB^{-\alpha/(\alpha+1)}$

See next column for why this prevents adaptation

A closer look at the lower bound

$R_T(\alpha) := \sup_{f\ \alpha\text{-Hölder}} R_T$ **Assume** $\forall \alpha, T \quad R_T(\alpha) \leqslant cT^{\theta(\alpha)}$

Lower bound yields: $R_T(\alpha) \geqslant cTR_T(\gamma)^{-\alpha/(\alpha+1)}$ **when** $\alpha \leqslant \gamma$

$$\forall \alpha \leqslant \gamma, \quad \theta(\alpha) \geqslant 1 - \theta(\gamma)\frac{\alpha}{\alpha+1} \qquad (\bigstar)$$

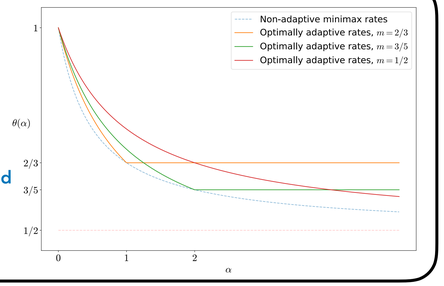
Minimal solutions to $(\bigstar)$

$\theta_m : \alpha \mapsto \max\left(m, 1 - m\frac{\alpha}{\alpha+1}\right)$

with $m \in [1/2, 1]$

No rates smaller than these can be reached

and these rate functions are everywhere above the usual rates



Matching the lower bound(s): 3 ingredients

Usual methods zoom in on promising regions, in a way that crucially depends on the regularity

e.g.
- Bubeck, Munos, Stoltz, Szepesvári '11 "X-Armed Bandits"
- Kleinberg, Slivkins, Upfal '11 "Bandits in metric spaces"
- Bull '15 "Adaptive-treed bandits", and many more (see full paper)

We do the opposite and zoom out

Discretize: Split the time budget into epochs; use a new discretization in each epoch.

Zoom Out: At each new epoch, reset the algorithm and start over a new regime of length double the previous one and with half fewer discrete arms.

Memorize past actions: Allow the discrete algorithm to pick an action uniformly among the actions played in each of the past epochs.

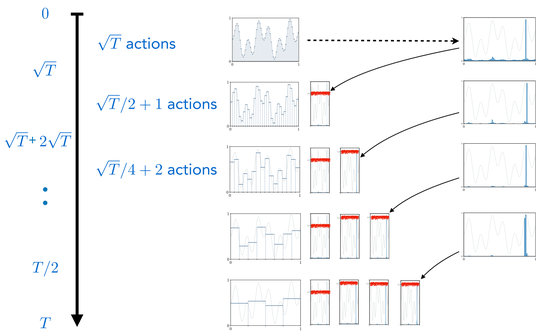
Algorithm: MeDZO

Set $K_i \approx 2^{-i}\sqrt{T}$; $T_i \approx 2^i\sqrt{T}$

For epochs 1 to $p \approx \log\sqrt{T}$

- For T_i rounds, run CAB1 with K_i-discretization and the memorized actions

Distributions of plays



Regret analysis

Denote by $R(j)$ the regret suffered during j-th epoch

After epoch j, whenever the discrete algorithm picks the memorized action from epoch j, the instantaneous regret suffered is smaller than

$$\max f - \underbrace{\frac{1}{T_j}\mathbb{E}\left[\sum_{\text{epoch } j} f(X_t)\right]}_{:= M(j),\ \text{the expected payoff of action memorized from epoch j}} = \frac{R(j)}{T_j}$$

Then by the guarantees of CAB1, for j < i $\quad R(i) \leqslant T_i\left(\max f - M(j)\right) + c\sqrt{T_i K_i}$

Even though we zoom out, the approximation error from the discretization does not grow too fast, thanks to the memorized actions.

$$R(i) \leqslant T_i\frac{R(j)}{T_j} + c\sqrt{T} \quad \text{for all } j < i$$

Sum over i and use the Hölder property for the early discretizations to get the regret bound

Replace $\sqrt{T}$ by T^m to obtain any rate among the θ_m's

Without the knowledge of α $\quad R_T \leqslant c_T \max\left(T^m, T^{1-m\alpha/(\alpha+1)}\right)$

Additional References

Bandits:
- Auer et. al '02 "Using confidence bounds for exploration-exploitation tradeoffs"

Adaptativity for simple regret/Optimisation:
- Grill, Valko, Munos '15 "Black-box optimization of noisy functions with unknown smoothness"
- Bartlett, Gabillon, Valko '19 "A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption"