Simple regret for infinitely many armed bandits

ALEXANDRA CARPENTIER and MICHAL VALKO
a.carpentier@statslab.cam.ac.uk and michal.valko@inria.fr

UNIVERSITY OF CAMBRIDGE
informatics mathematics Inria

Setting

$\mathbb{A}_t = \{\nu_1, \dots, \nu_{K_t}\}$

At time t ≤ n

pull a known arm k_t or try a new one

$K_{t+1} = K_t$ $\qquad$ $K_{t+1} = K_t + 1$

$\mathbb{A}_{t+1} = \mathbb{A}_t$ $\qquad$ $\mathbb{A}_{t+1} = \mathbb{A}_t \cup \{\nu_{K_t+1}\}$

$\nu_{K_t+1} \sim \tilde{\mathcal{L}}$

get the sample $X_t \sim \nu_{k_t}$

Reward only at the end

$$r_n = \bar{\mu}^* - \mu_{\hat{k}}$$

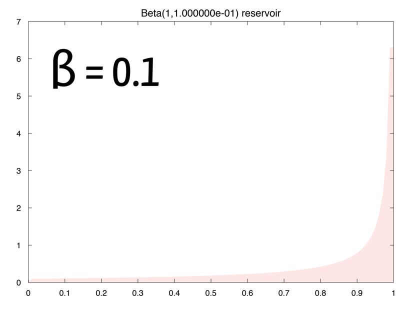
arm selection tradeoff

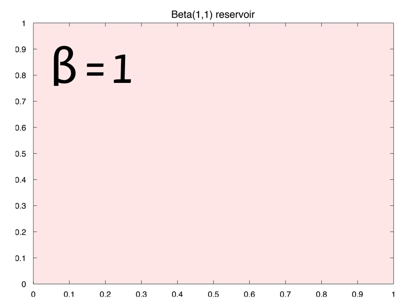
- take many to get at least one good
- take few to evaluate them well

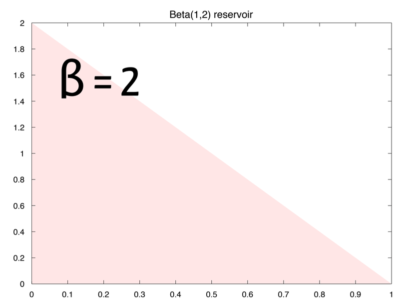
different from exploration/exploitation tradeoff

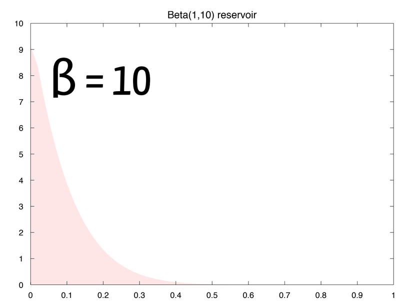
Reservoirs

$\mathbb{P}_{\mu \sim \mathcal{L}}(\bar{\mu}^* - \mu \geq \varepsilon) \approx \varepsilon^\beta$









Definitions

$$b = \min(\beta, 2)$$

$$\bar{T}_\beta = \lceil A(n) n^{b/2} \rceil$$

$$A(n) = \begin{cases} A, & \text{if } \beta < 2 \\ A/\log(n)^2, & \text{if } \beta = 2 \\ A/\log(n), & \text{if } \beta > 2 \end{cases}$$

$$\widehat{\mu}_{k,t} = \frac{1}{T_{k,t}} \sum_{u=1}^{T_{k,t}} X_{k,u}$$

$$\bar{t}_\beta = \lfloor \log_2(\bar{T}_\beta) \rfloor$$

Algorithm

SiRI - Simple Regret for Infinitely Many Armed Bandits

START: Sample $\bar{T}_b$ Arms and pull each once

Update B-values (estimates + confidence intervals)

$$B_{k,t} \leftarrow \widehat{\mu}_{k,t} + 2\sqrt{\frac{C}{T_{k,t}} \log\left(2^{2\bar{t}_\beta/b}/(T_{k,t}\delta)\right)} + \frac{2C}{T_{k,t}} \log\left(2^{2\bar{t}_\beta/b}/(T_{k,t}\delta)\right)$$

Pick arm k_t with the highest B value

Pull arm k_t to double the samples from it

END: return the arm most pulled

ANR ExtraLearn

Where is it useful?

- When we are faced with many choices
  - but we **can't try them all** even once.
- Applicable to finite but extremely large cases.
- Single feature selection (biomarkers).

Other infinite bandits

- X-armed bandits, bandits in metric spaces, ...
- linear bandits, convex bandits, ...

All require contextual information (embedding).

References

prior work that considered the cumulative regret case
- Berry et al. 1997
  - formalization and motivation
  - **asymptotic** result
- Wang et al. 2008 - UCB-F
  - **finite** time result
- Bonald and Proutière, 2013
  - tight results for the **uniform** reservoir

simple regret work that considered the finite arm case
- Jamieson et al. 2014 - lil'UCB
  - best arm in the fixed **confidence** setting
- Audibert et al. 2010 - UCB-E
  - best arm in the fixed **budget** setting

Unknown β?

Solution: **$\bar{\beta}$-SiRI** algorithm

- Devote $n^{1/2}$ samples to estimate β.
- Get $n^{1/4}$ arms and sample them $n^{1/4}$ times each.
- Same guarantees as for SiRI (up to loglog n).

Rewards in [0,1] with $\mu^*=1$?

The variance of the near-optimal arms is small.

Empirical Bernstein-modified SiRI (idea by Wang et al. 2008)

Improved minimax optimal rates (up to polylog n)

β≤1: whp $\mathcal{O}\left(\frac{1}{n}\,\text{polylog}\, n\right)$

β>1: whp $\mathcal{O}\left(\left(\frac{1}{n}\right)^{1/\beta}\text{polylog}\, n\right)$

Anytime algorithm?

2 options

1) doubling trick

2) UCB-AIR method (Wang et at. 2008)

In both cases: regret only worsened by polylog n.

Comparison

	minimax rates	phase transition
cumulative regret	$\max(n^{\beta/(\beta+1)}, n^{1/2})$	β = 1
cumulative regret bounded	$n^{\beta/(\beta+1)}$	none
simple regret	$\max(n^{-1/\beta}, n^{-1/2})$	β = 2
simple regret bounded	$\max(n^{-1/\beta}, n^{-1})$	β = 1

Upper bounds of SiRI

β<2: whp $\quad r_n \leq E n^{-1/2}$

β>2: whp $\quad r_n \leq E(n \log n)^{-1/\beta}\, \text{polyloglog}\, n$

β=2: whp $\quad r_n \leq E n^{-1/2} \log n\, \text{polyloglog}\, n$

Lower bounds

β<2: wp > 1/3 $\quad \inf_{\mathcal{A}} \sup_{\tilde{\mathcal{L}} \in \mathcal{S}_\beta} r_n \geq v n^{-1/2}$

β≥2: wp > 1/3 $\quad \inf_{\mathcal{A}} \sup_{\tilde{\mathcal{L}} \in \mathcal{S}_\beta} r_n \geq v n^{-1/\beta}$

Proof sketch

Based on 2 events that hold with high probability:

ξ_1 - controls the number arm at a given distance from μ^*

ξ_2 - controls the distance between empirical and true means μ

Given ξ_1 and ξ_2 we show that:

- Given the suboptimality gap we can bound the number of suboptimal arms.
- Among $\bar{T}_b$ arms pulled, there is at least one good enough.
- Empirical means are close to the true ones. (True means are random!)
- We can bound the number of suboptimal arms.
- We can upper bound the number of suboptimal pulls.
- There is a near-optimal arm pulled more than n/2 times.
- By definition, this near-optimal arm is selected by SiRI.

Experiments