

Multi-Armed Bandits

Florian Turati

A Classical Dilemma

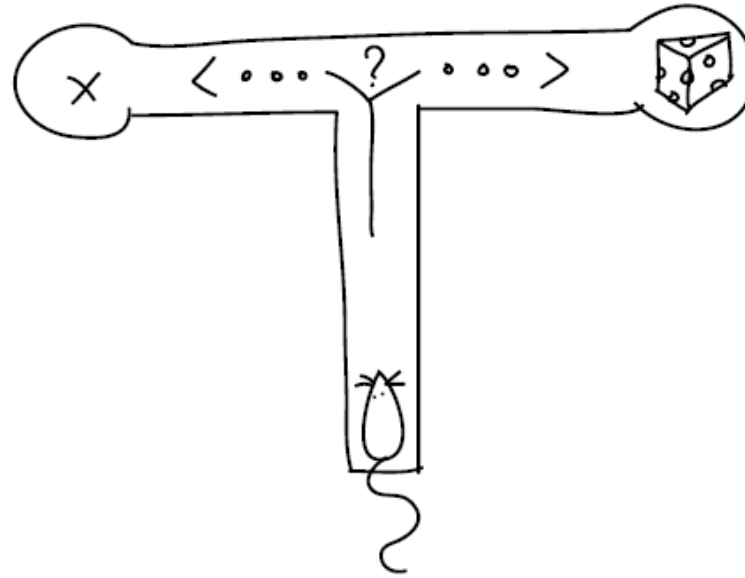


Round	1	2	3	4	5	6	7	8	9	10
Left	0		10	0		0				10
Right		10			0		0	0	0	

Outline

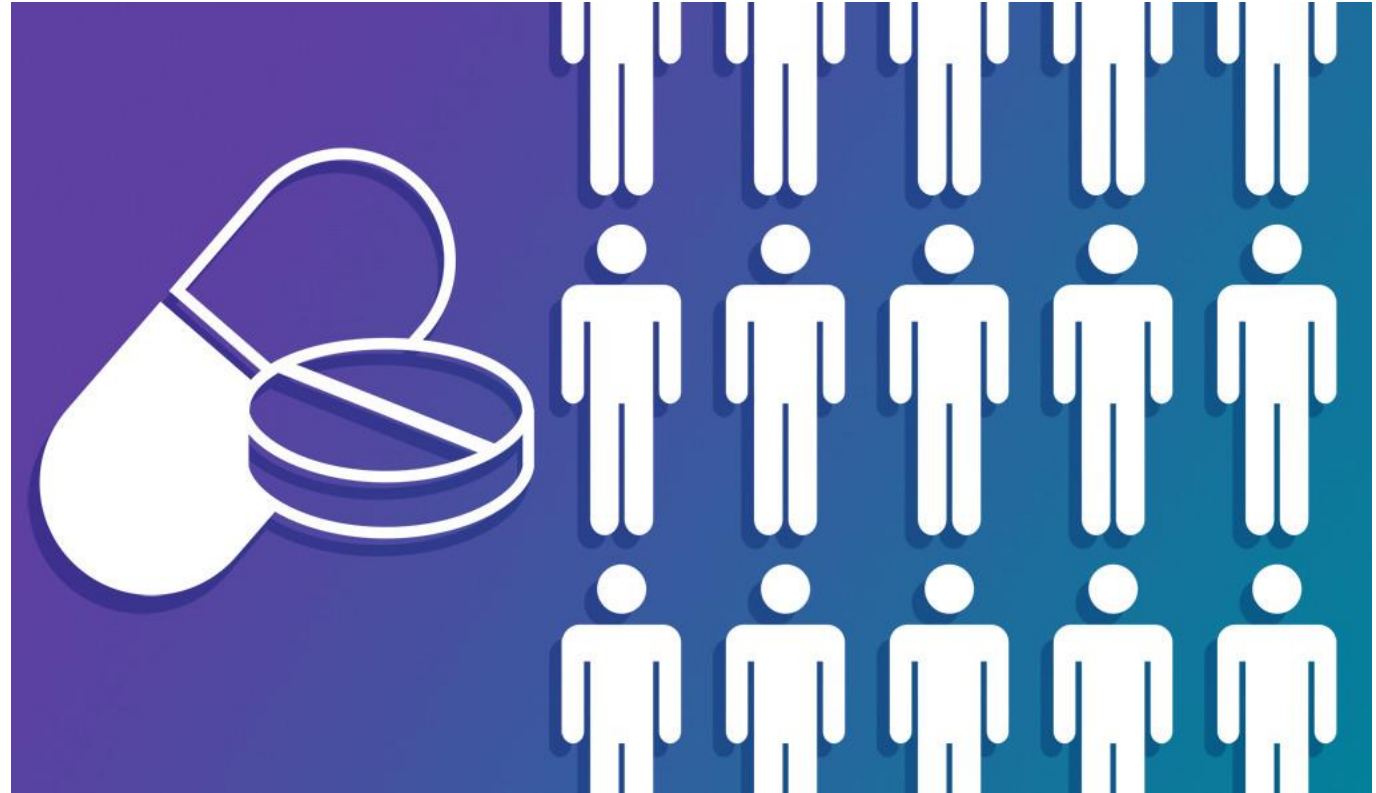
- **Motivation**
- Stochastic bandits
- Bayesian bandits
- Lipschitz bandits

What's in the name ? A brief history



Applications

- Clinical trials



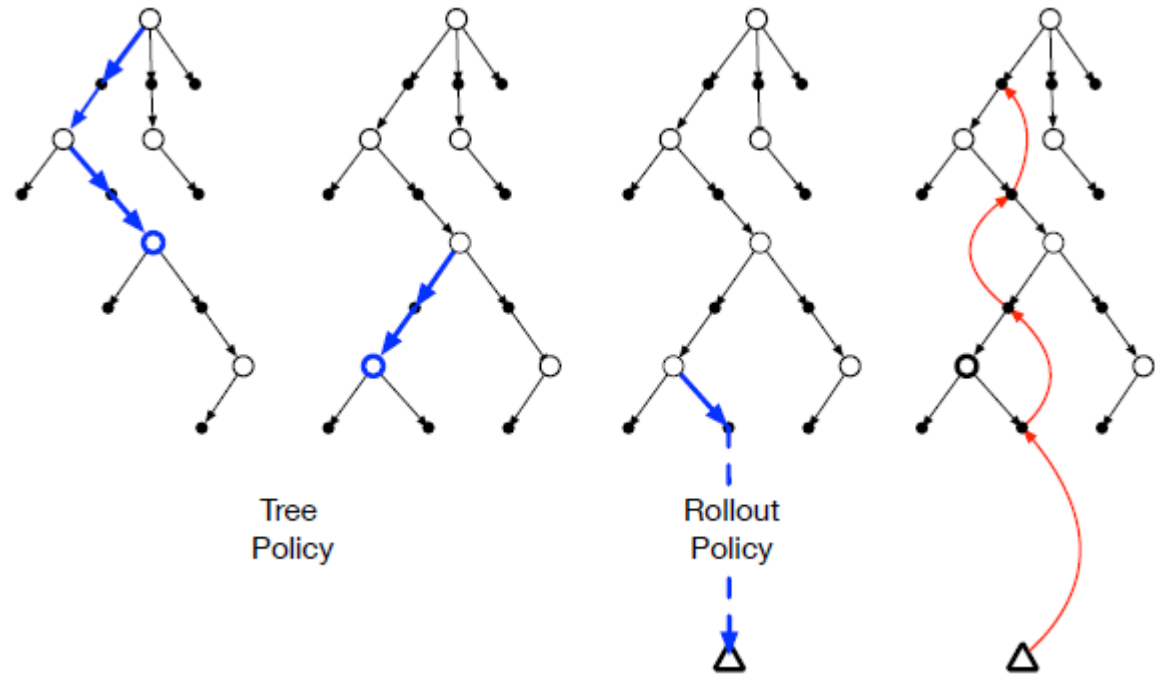
Applications

- Web interfaces
- Ad placement
- Recommender systems

The screenshot shows the CNN website interface. At the top, there is a travel advertisement for CNN Travel with the text "A journey of a thousand steps begins with a single click" and a "START MY JOURNEY" button. Below this is a navigation bar with categories like World, US Politics, Business, Health, Entertainment, Style, Travel, Sports, and Videos. A prominent red banner at the top of the main content area reads "TRUMP PREVIEWS REOPENING PLANS" and "US President told governors they'll call the shots in their states as he shared guidelines to reopen in phases starting May 1". Below the banner, there are navigation links for "CORONAVIRUS" (Live updates, All coronavirus stories, Global virus map, Key symptoms, Podcast, Newsletter) and "TRENDING" (Edward Snowden). The main content area features a large article titled "White House gives guidelines on reopening economy" with a video player showing President Trump speaking. To the right, there are several smaller article thumbnails, including "What Trump could learn from Angela Merkel about dealing with coronavirus" and "Ivanka and Jared Kushner don't think the coronavirus rules apply to them". A sidebar on the right contains a list of trending news items such as "Images of Venice from space show how coronavirus has changed city's iconic canals" and "Bolsonaro fires popular health minister".

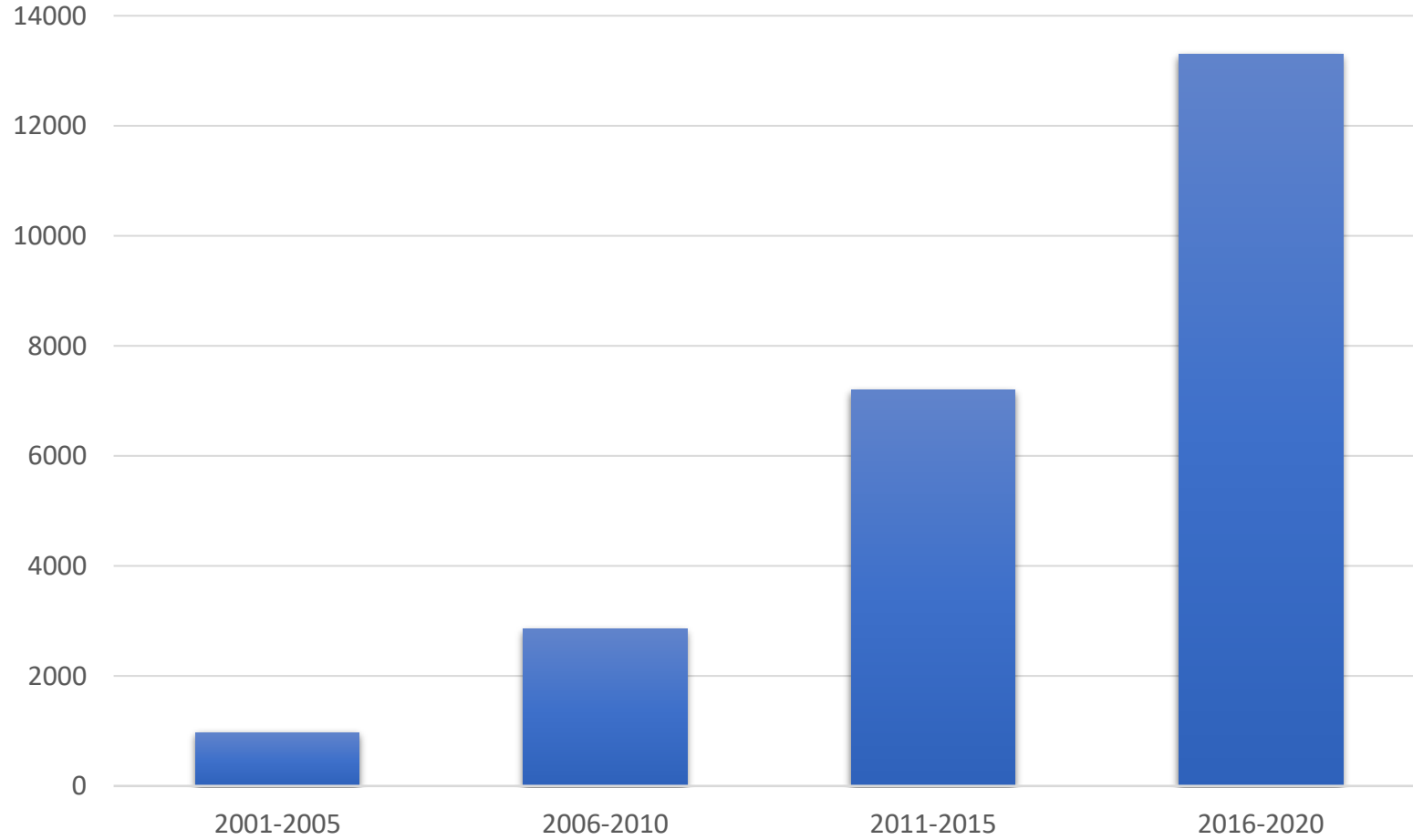
Applications

- Game tree search



Research

Number of papers



Outline

- Motivation
- **Stochastic bandits**
- Bayesian bandits
- Lipschitz bandits

A Classical Dilemma



Round	11	12	13	14	15	16	17	18	19	20
Left	10		0		10	0	0	10	0	10
Right		0		10						

Stochastic bandits (MAB with IID rewards)

Basic protocol :

Given K arms, T rounds.

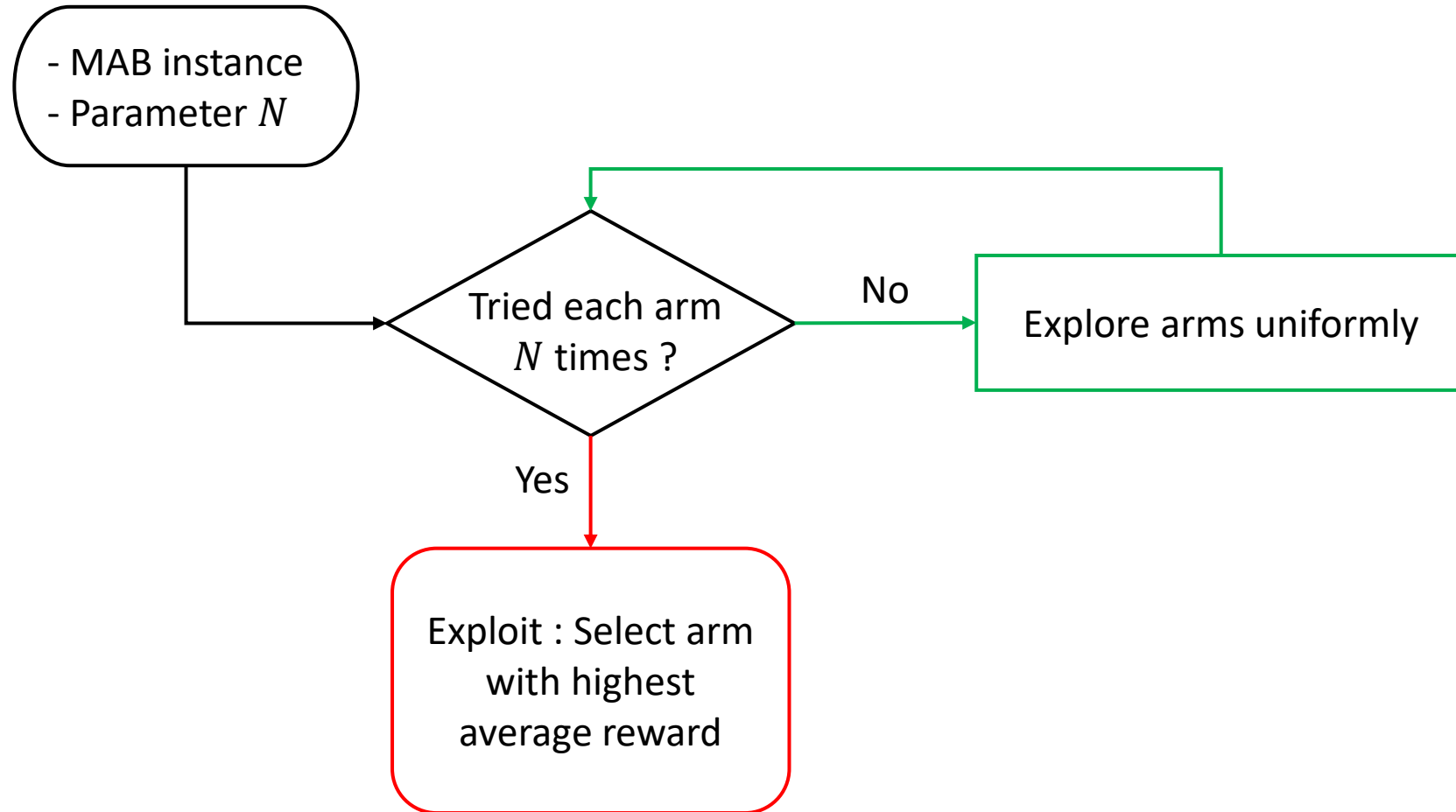
For each round $t \leq T$

1. Algorithm picks arm a_t
2. Algorithm observes reward r_t for chosen arm

Goal: maximize total reward over T rounds

Primary interest: mean reward vector μ ,
where $\mu(a) = \mathbb{E}[\mathbb{D}_a]$ is the mean reward of arm a .

Uniform Exploration



Regret

$$R(T) = \mu^* \cdot T - \sum_{t=1}^T \mu(a_t)$$

with $\mu^* := \max_{a \in A} \mu(a)$ and $\mu(a) := \mathbb{E}[\mathbb{D}_a]$

Regret Analysis

Hoeffding inequality (clean event):

$$\Pr\{|\bar{\mu}(a) - \mu(a)| \leq r(a)\} \geq 1 - \frac{2}{T^4}$$

with confidence radius $r(a) = \sqrt{\frac{2 \log T}{N}}$

Regret Analysis

case $K = 2$ arms

$$\mu(a) + r(a) \geq \bar{\mu}(a) > \bar{\mu}(a^*) \geq \mu(a^*) - r(a^*)$$

$$\bar{\mu}(a^*) - \bar{\mu}(a) \leq r(a) + r(a^*) = o\left(\sqrt{\frac{\log T}{N}}\right)$$

Regret Analysis

$$R(T) \leq N + O\left(\sqrt{\frac{\log T}{N}} \times (T - 2N)\right) \leq N + O\left(\sqrt{\frac{\log T}{N}} \times T\right)$$

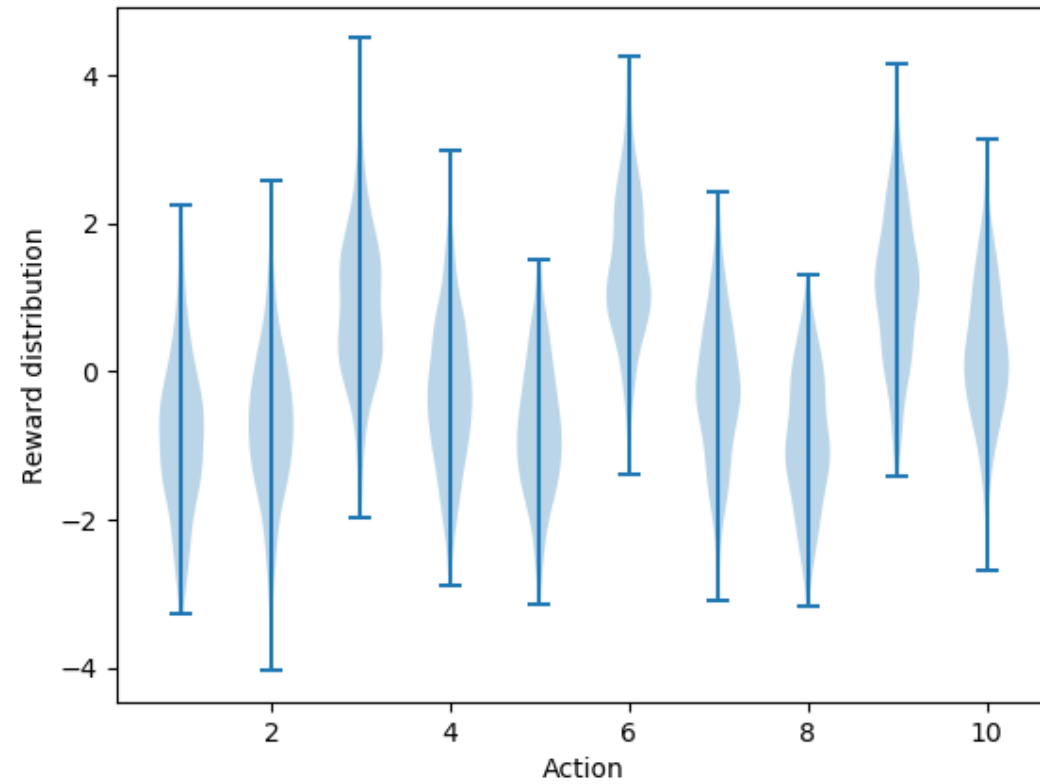
minimize with $N = T^{\frac{2}{3}}(\log T)^{\frac{1}{3}}$

$$R(T) \leq O\left(T^{\frac{2}{3}}(\log T)^{\frac{1}{3}}\right)$$

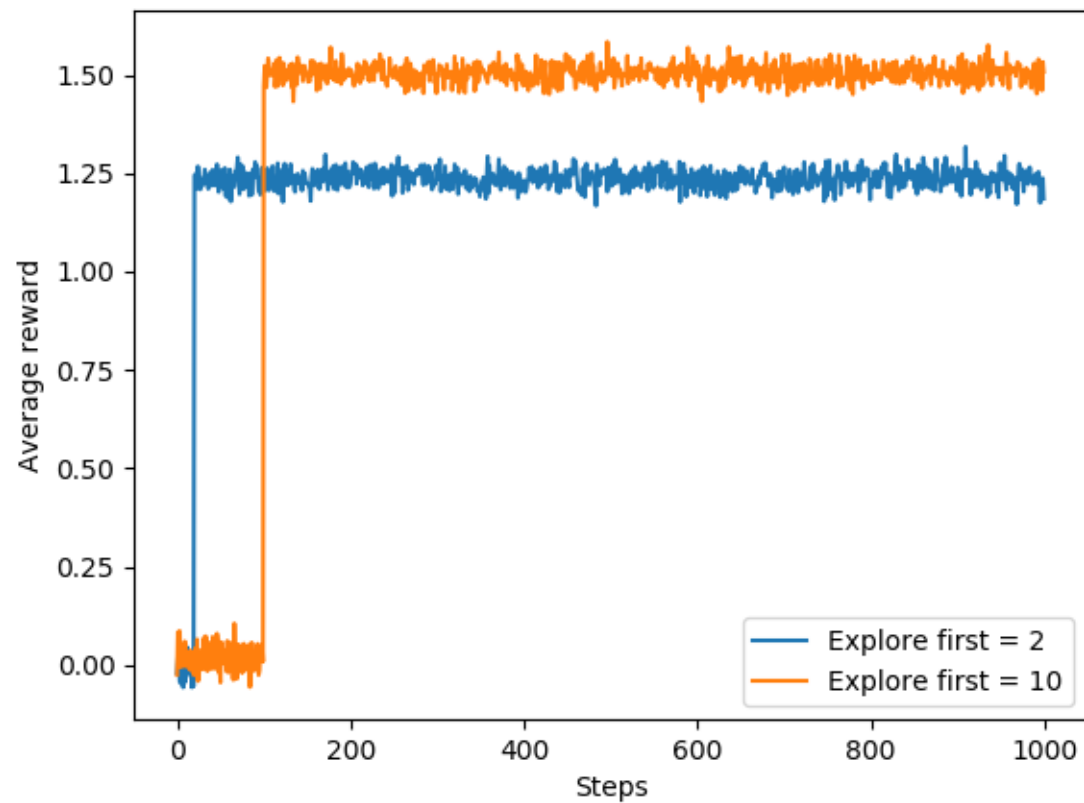
Regret Analysis

$$\mathbb{E}[R(T)] \leq T^{\frac{2}{3}} \times O(K \log T)^{\frac{1}{3}}$$

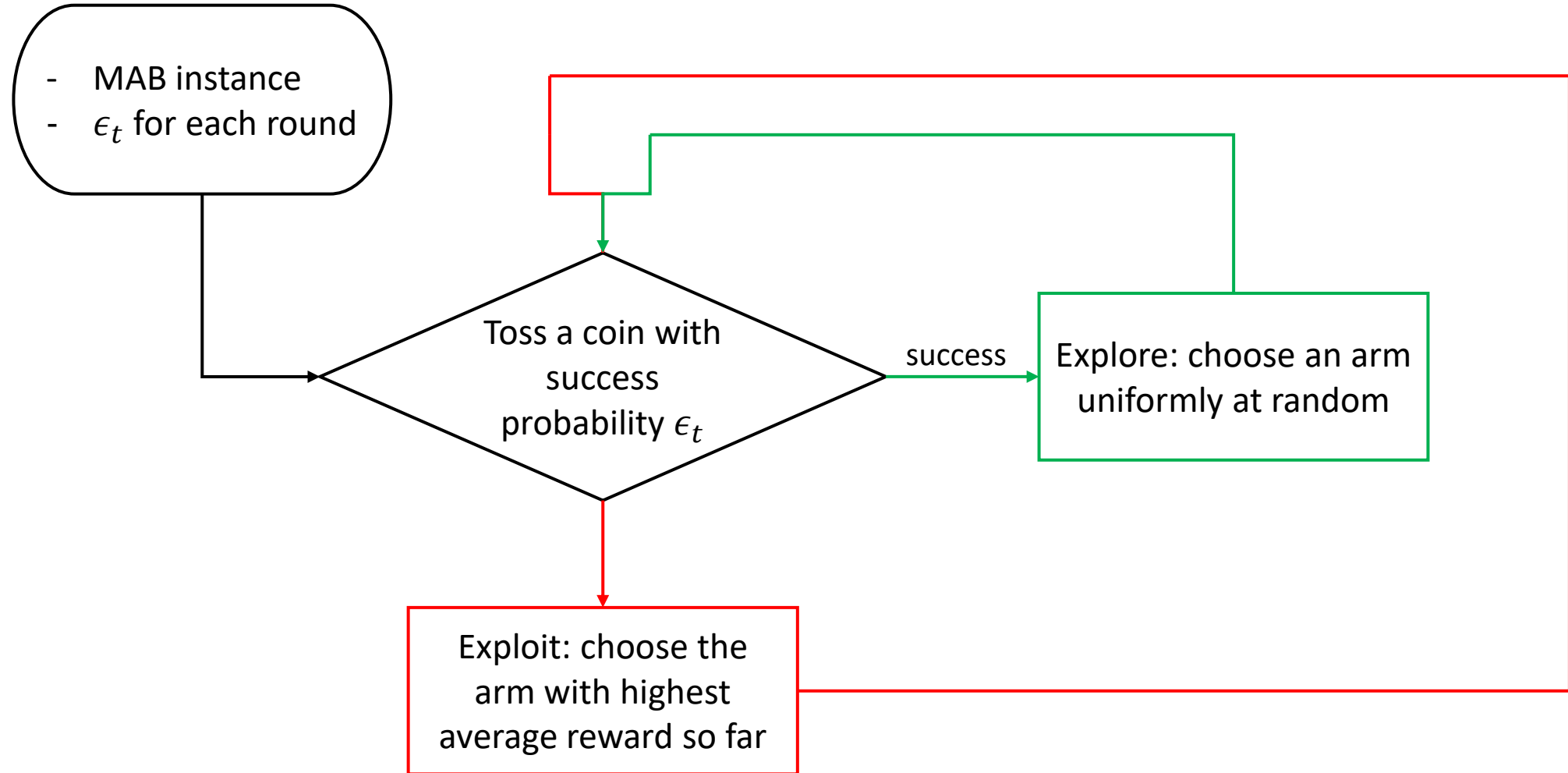
10-armed Testbed



Benchmark: 10-armed Testbed



ϵ -greedy

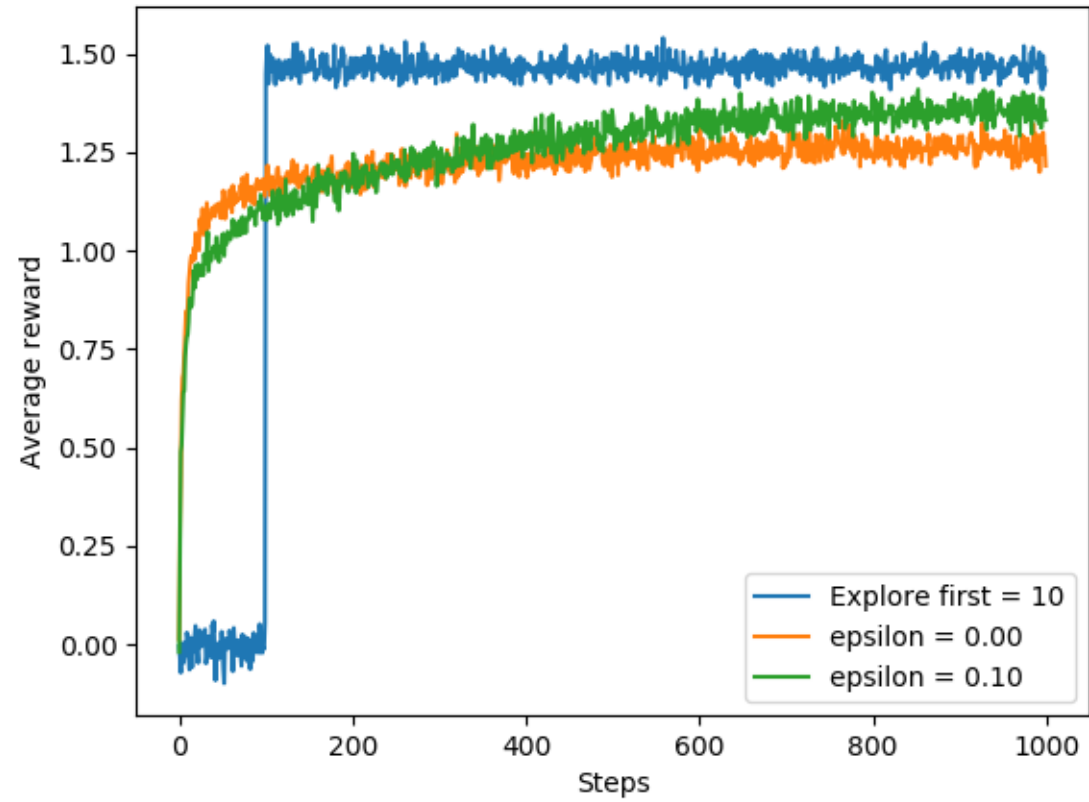


Regret Analysis

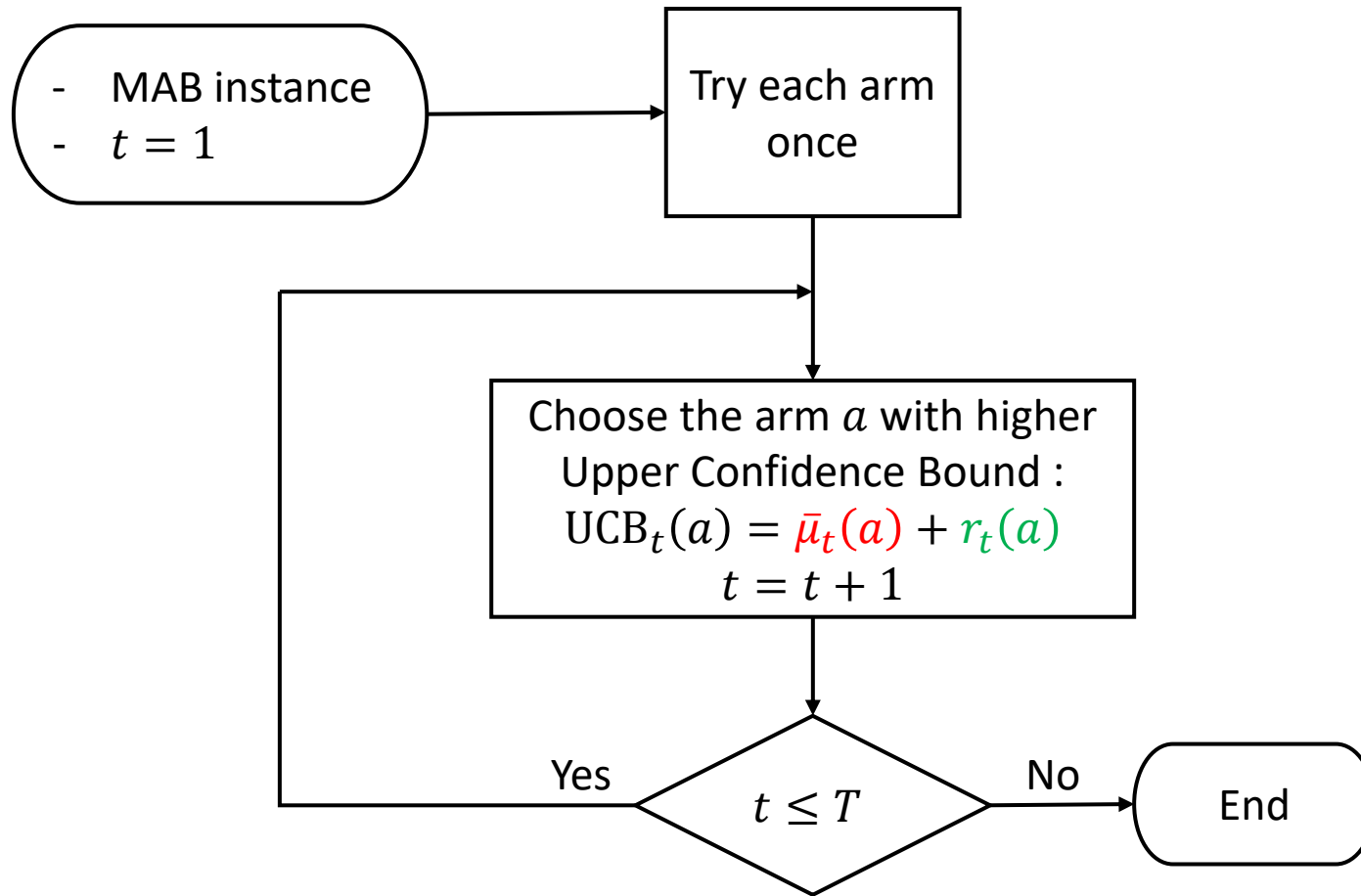
With exploration probabilities $\epsilon_t = t^{-\frac{1}{3}} \cdot (K \log t)^{\frac{1}{3}}$
For each round t :

$$\mathbb{E}[R(t)] \leq t^{\frac{2}{3}} \times O(K \log t)^{\frac{1}{3}}$$

Benchmark



Upper Confidence Bound exploration



Upper Confidence Bound exploration

UCB1:

Try each arm once:

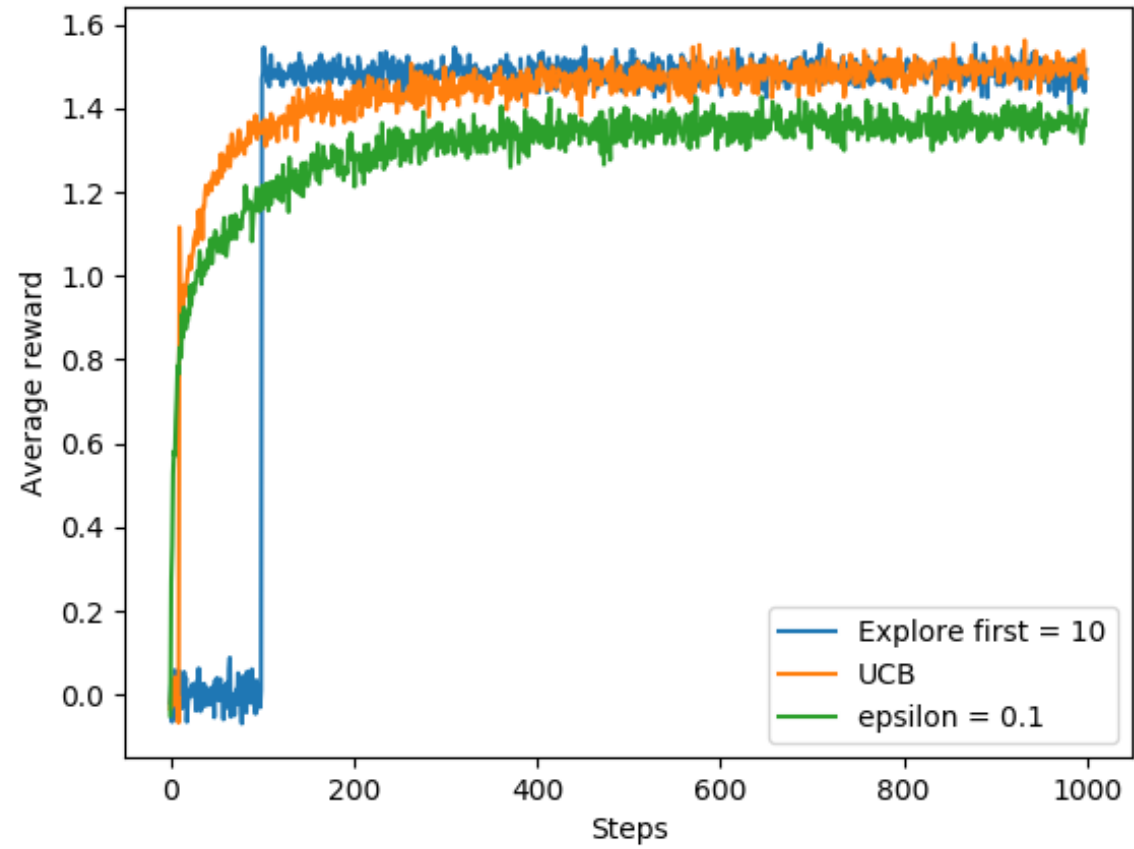
In each round t , pick $\operatorname{argmax}_{a \in A} UCB_t(a)$, where $UCB_t(a) = \bar{\mu}_t(a) + r_t(a)$

Recall: confidence radius $r_t(a) = \sqrt{\frac{2 \log t}{n_t(a)}}$

Regret Analysis

$$\mathbb{E}[R(t)] \leq O(\sqrt{Kt \log T}) \text{ for all rounds } t \leq T$$

Benchmark



Outline

- Motivation
- Stochastic bandits
- **Bayesian bandits**
- Lipschitz bandits

A Classical Dilemma



Round	1	2	3	4	5	6	7	8	9	10
Left	0		1	0		0				1
Right		1			0		0	0	0	

Bayesian Bandits

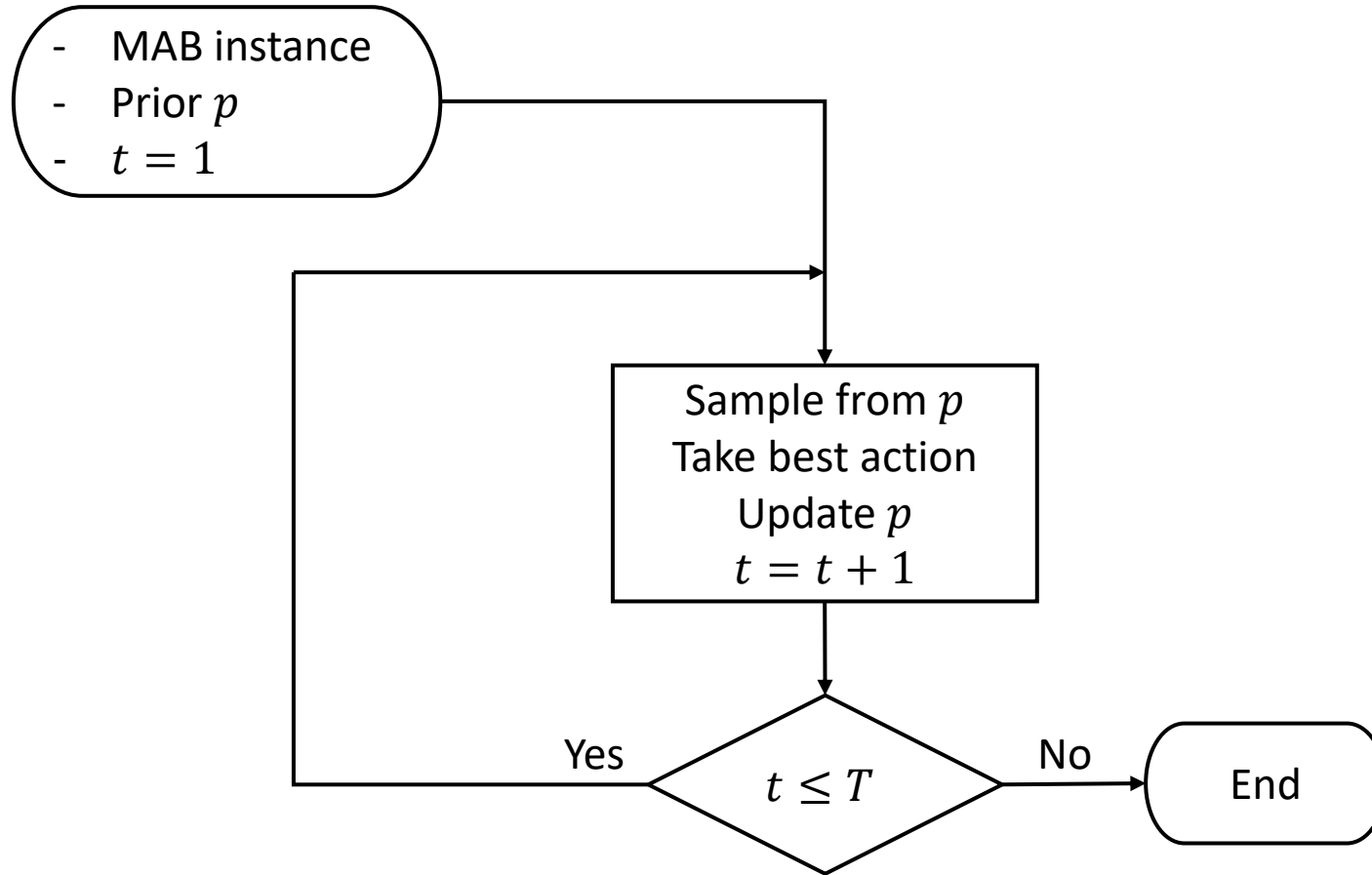
Bayesian assumption : $I \sim \mathbb{P}$

$$\mu(a) = \mathbb{E}[\mathbb{D}_a]$$

Bayesian regret:

$$BR(T) := \mathbb{E}_{I \sim \mathbb{P}}[\mathbb{E}[R(T)|I]] = \mathbb{E}_{I \sim \mathbb{P}}[\mu^* \cdot T - \sum_{t \in [T]} \mu(a_t)]$$

Thompson Sampling



Terminology

t -history:

$$H_t = ((a_1, r_1), \dots, (a_t, r_t)) \in (A \times \mathbb{R})^t$$

feasible t -history:

$$H = ((a'_1, r'_1), \dots, (a'_t, r'_t)) \in (A \times \mathbb{R})^t$$

with $\Pr[H_t = H] > 0$

Thompson Sampling

For each round t

observe $H_{t-1} = H$, for some feasible $(t - 1)$ -history H :

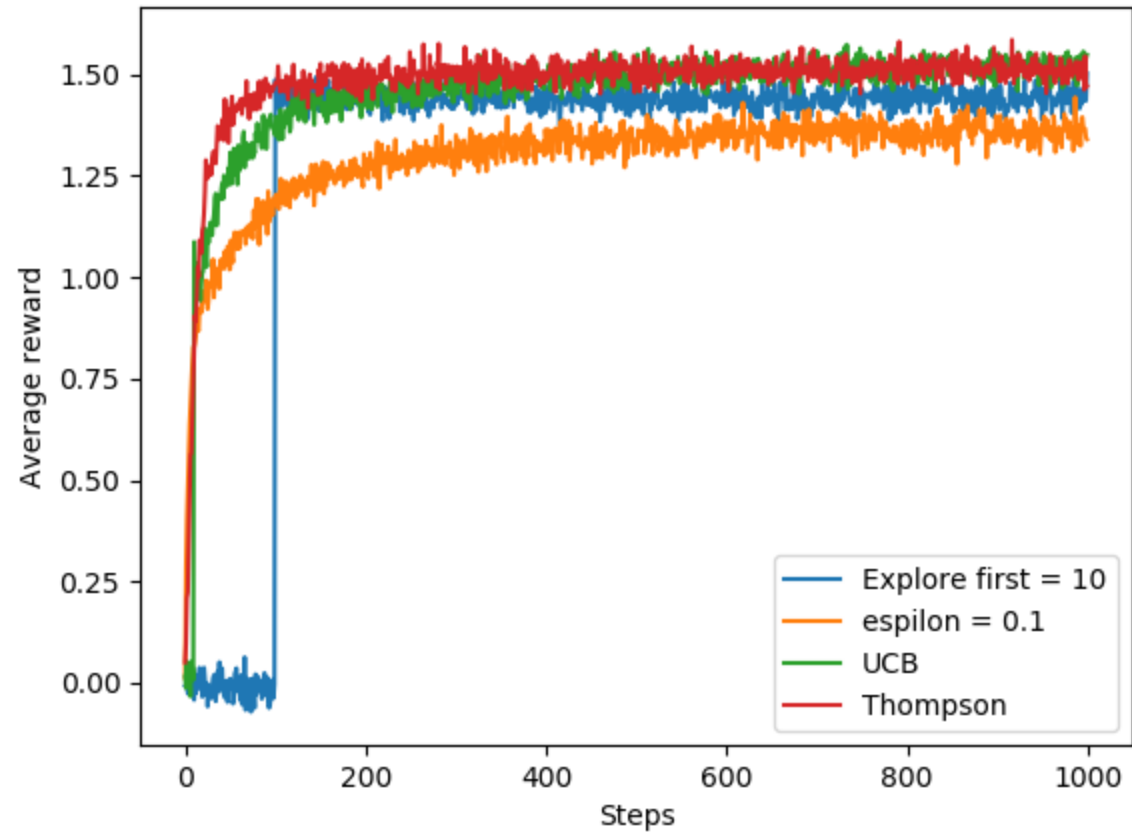
Draw arm a_t independently from distribution $p_t(\cdot | H)$, where

$$p_t(a|H) := \Pr[a^* = a | H_{t-1} = H] \quad \text{for each arm } a$$

Bayesian regret analysis

$$BR(T) = O(\sqrt{KT \log(T)})$$

Benchmark



Outline

- Motivation
- Stochastic bandits
- Bayesian bandits
- **Lipschitz bandits**

A Classical Dilemma



Round	1	2	3	4	5	6	7	8	9	10
Left	0				10	0				10
Right		10	0	0			10	0	0	

Continuum-armed bandits (CAB)

Lipschitz condition:

$$|\mu(x) - \mu(y)| \leq L \cdot |x - y| \text{ for any two arms } x, y \in X = [0,1]$$

Fixed discretization

Discretization:

Finite set of arms $S \subset X$

Best arm in S :

$$\mu^*(S) = \sup_{x \in S} \mu(x)$$

Discretization error:

$$\text{DE}(S) = \mu^*(X) - \mu^*(S)$$

Regret Analysis

$$\begin{aligned}\mathbb{E}[R(T)] &= T \cdot \mu^*(X) - W(ALG) \\ &= (T \cdot \mu^*(S) - W(ALG)) + T \cdot (\mu^*(X) - \mu^*(S)) \\ &= R_S(T) + T \cdot \text{DE}(S)\end{aligned}$$

where $W(ALG)$ is the total reward of the algorithm

$$\mathbb{E}[R(T)] \leq O\left(\sqrt{|S|T \log T}\right) + T \cdot \text{DE}(S)$$

Regret Analysis

Fixed uniform discretization :

Consider $S \subset X = [0,1]$, $|S| = \left\lceil \frac{1}{\epsilon} \right\rceil$

$$\text{DE}(S) \leq L \cdot \epsilon$$

$$\mathbb{E}[R(T)] \leq O\left(L^{\frac{1}{3}} \cdot T^{\frac{2}{3}} \cdot \log^{\frac{1}{3}}(T)\right)$$

Lipschitz MAB

Recall CAB:

$$|\mu(x) - \mu(y)| \leq L \cdot |x - y| \text{ for any two arms } x, y \in X = [0,1]$$

Now:

$$|\mu(x) - \mu(y)| \leq \mathcal{D}(x, y) \text{ for any two arms } x, y$$

Regret Analysis

metric space: $X = [0,1]^d$ under l_p metric ($p \geq 1$)

Consider $S \subset X$, $|S| = \left(\left\lceil \frac{1}{\epsilon} \right\rceil\right)^d$

$$\text{DE}(S) \leq c_{p,d} \cdot \epsilon$$

Regret Analysis

Recall:

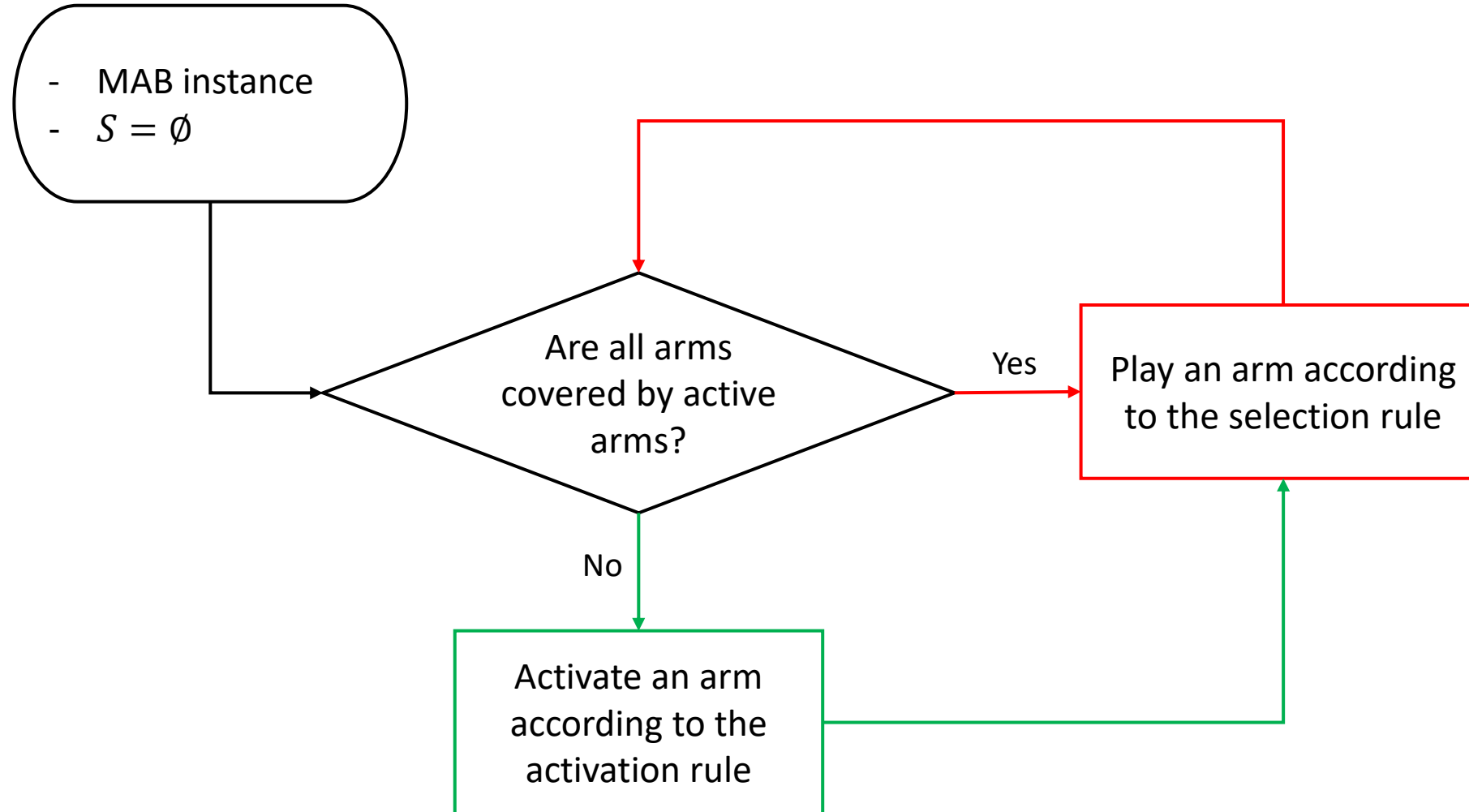
$$\mathbb{E}[R(T)] \leq O(\sqrt{|S|T \log T}) + T \cdot \text{DE}(S)$$

$$\mathbb{E}[R(T)] \leq O\left(T^{\frac{d+1}{d+2}} (c \log T)^{\frac{1}{d+2}}\right)$$

Adaptive discretization

$$\text{DE}(S) \leq \mathcal{D}(S, x^*) := \min_{x \in S} \mathcal{D}(x, x^*)$$

Zooming Algorithm



Zooming Algorithm

Initialize set of active arms $S \leftarrow \emptyset$

For each round t

if some arm y is not covered by confidence balls of active arms **then**

pick any such arm y and “activate” it: $S \leftarrow S \cup \{y\}$

play an active arm x with the largest $\text{index}_t(x)$

Activation rule

Selection rule

Zooming Algorithm: Notations

Confidence radius:

$$r_t(x) = \sqrt{\frac{2 \log T}{n_t(x)+1}}, \quad |\mu(x) - \mu_t(x)| \leq r_t(x)$$

Confidence Ball:

$$B_t(x) = \{y \in X: \mathcal{D}(x, y) \leq r_t(x)\}$$

Zooming Algorithm: Activation Rule

Suppose arm y not active and $\mathcal{D}(x, y) \ll r_t(x)$

invariant:

all arms are covered by confidence balls of the active arms

Activation rule:

If some arm y becomes uncovered by confidence balls of the active arms, activate y

Zooming Algorithm: Selection Rule

$$\text{index}_t(x) = \bar{\mu}_t(x) + 2r_t(x)$$

Recall UCB1:

$$UCB_t(x) = \bar{\mu}_t(x) + r_t(x)$$

Selection rule:

Play active arm with the largest index

Regret Analysis

$$\mathbb{E}[R(T)] \leq O\left(T^{\frac{d+1}{d+2}}(c \log T)^{\frac{1}{d+2}}\right)$$

Conclusion

Summary:

- Stochastic bandits
- Bayesian bandits
- Lipschitz bandits

Next:

Contextual bandits

References:

Introduction to Multi-Armed Bandits

– Aleksandrs Slivkins

Reinforcement Learning : An introduction (second edition)

– Richard S. Sutton, Andrew G. Barto

Bandit Algorithms

– Tor Lattimore, Csaba Szepesvári