Multi-Armed Bandits

Florian Turati

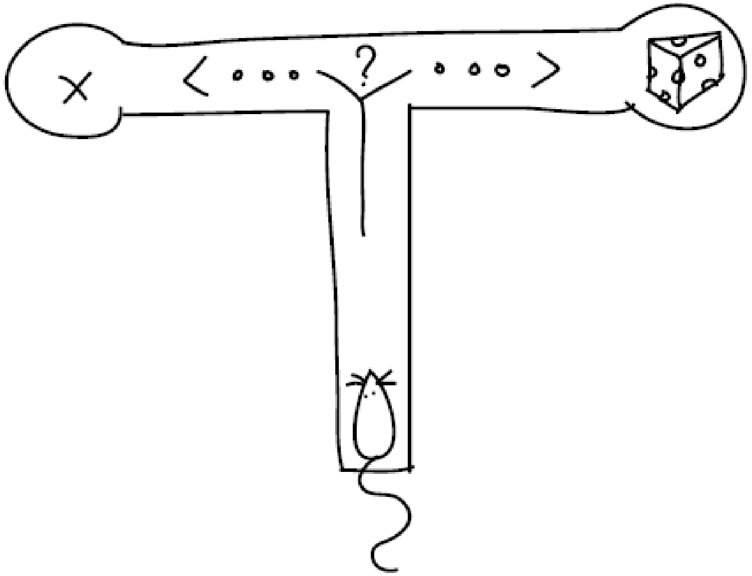
A Classical Dilemma

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

Outline

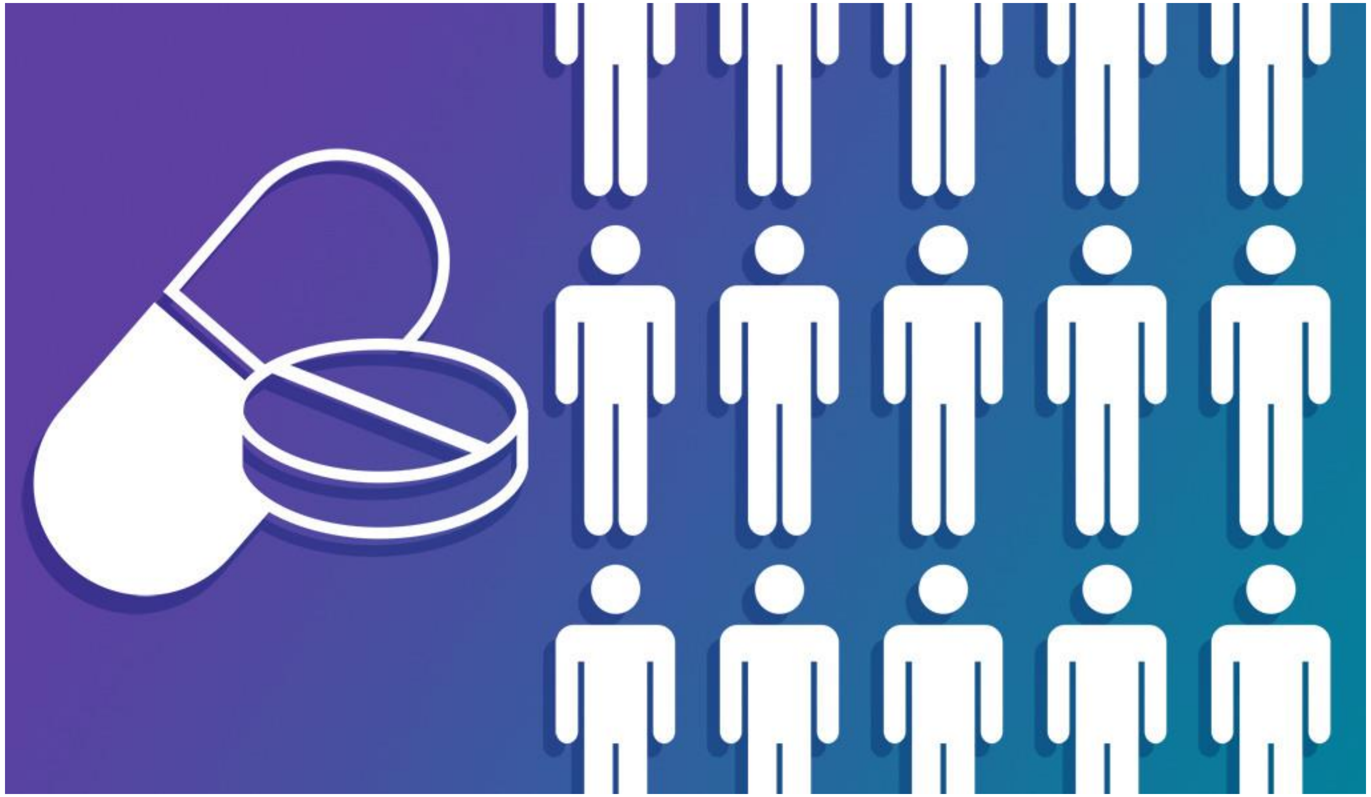
- **Motivation**
- Stochastic bandits
- Bayesian bandits
- Lipschitz bandits

What’s in the name ? A brief history



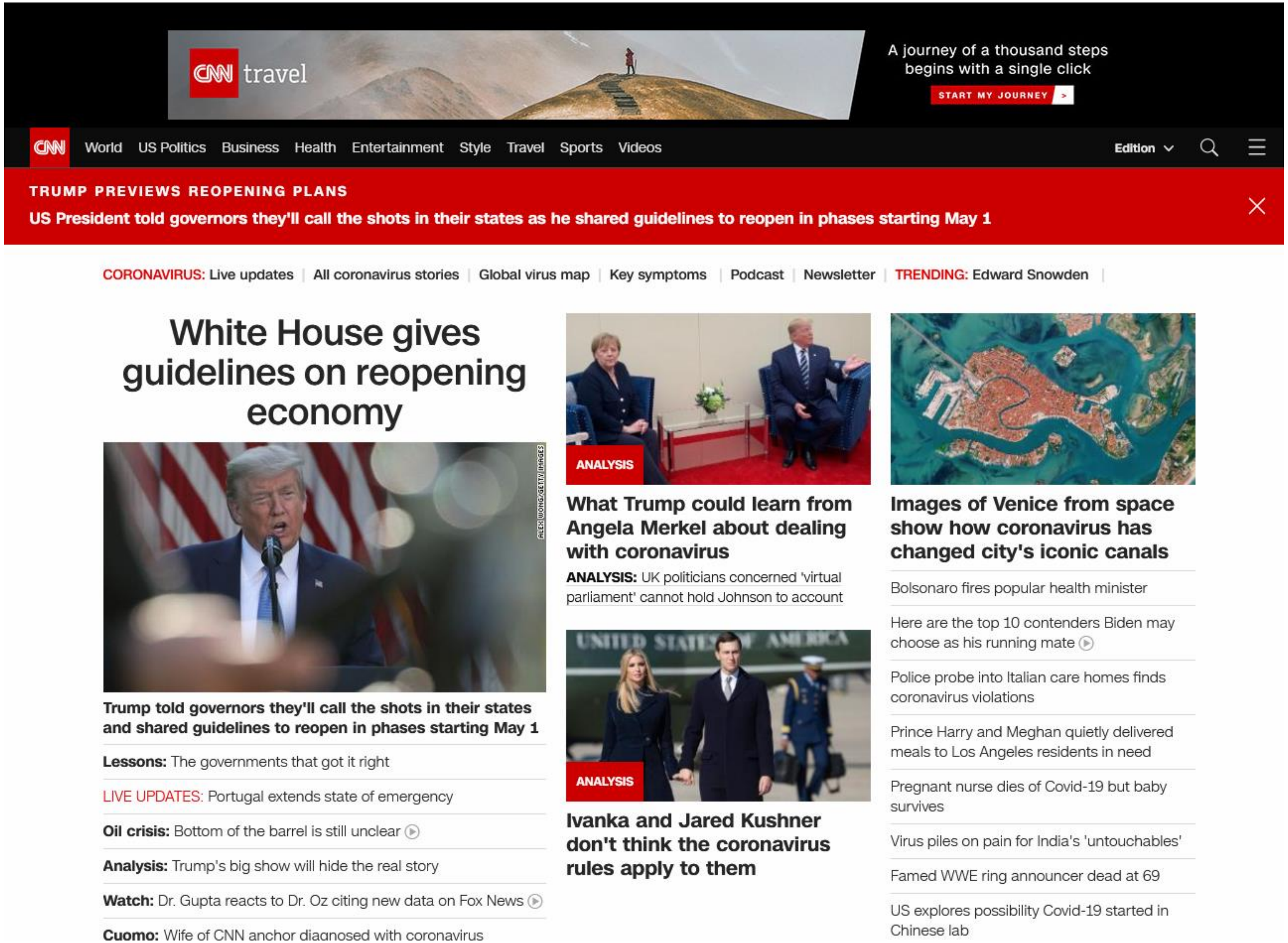
Applications

- Clinical trials



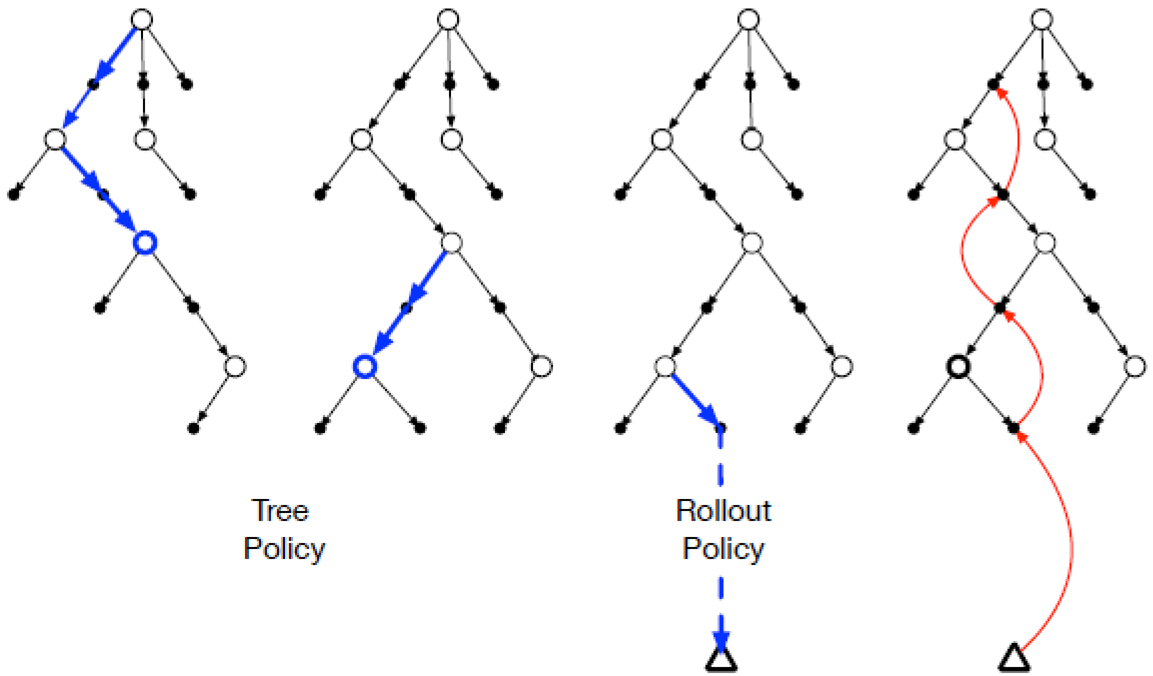
Applications

- Web interfaces
- Ad placement
- Recommender systems

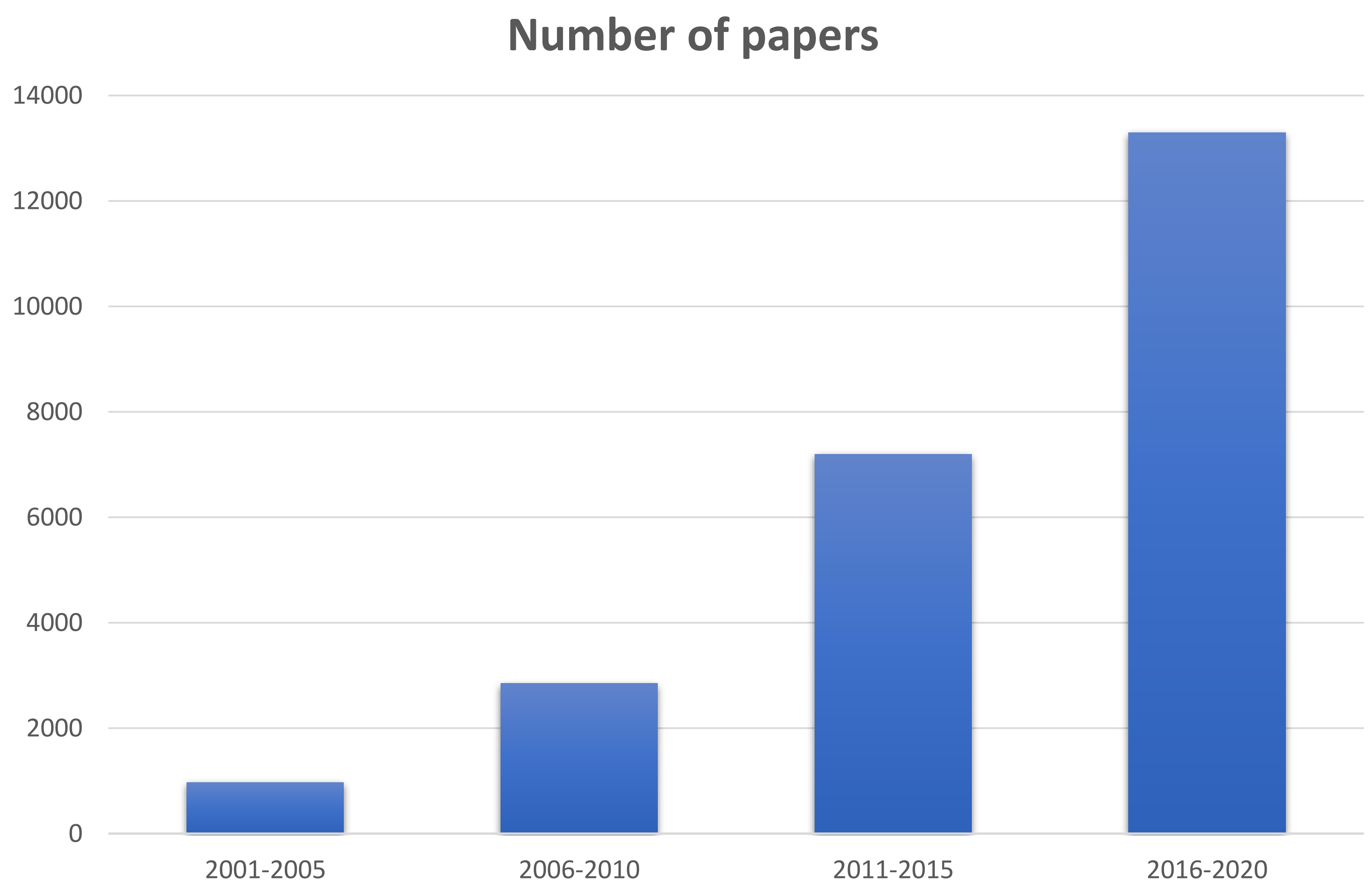


Applications

- Game tree search



Research


Number of papers
14000
12000
10000
8000
6000
4000
2000
0
2001-2005
2006-2010
2011-2015
2016-2020

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)

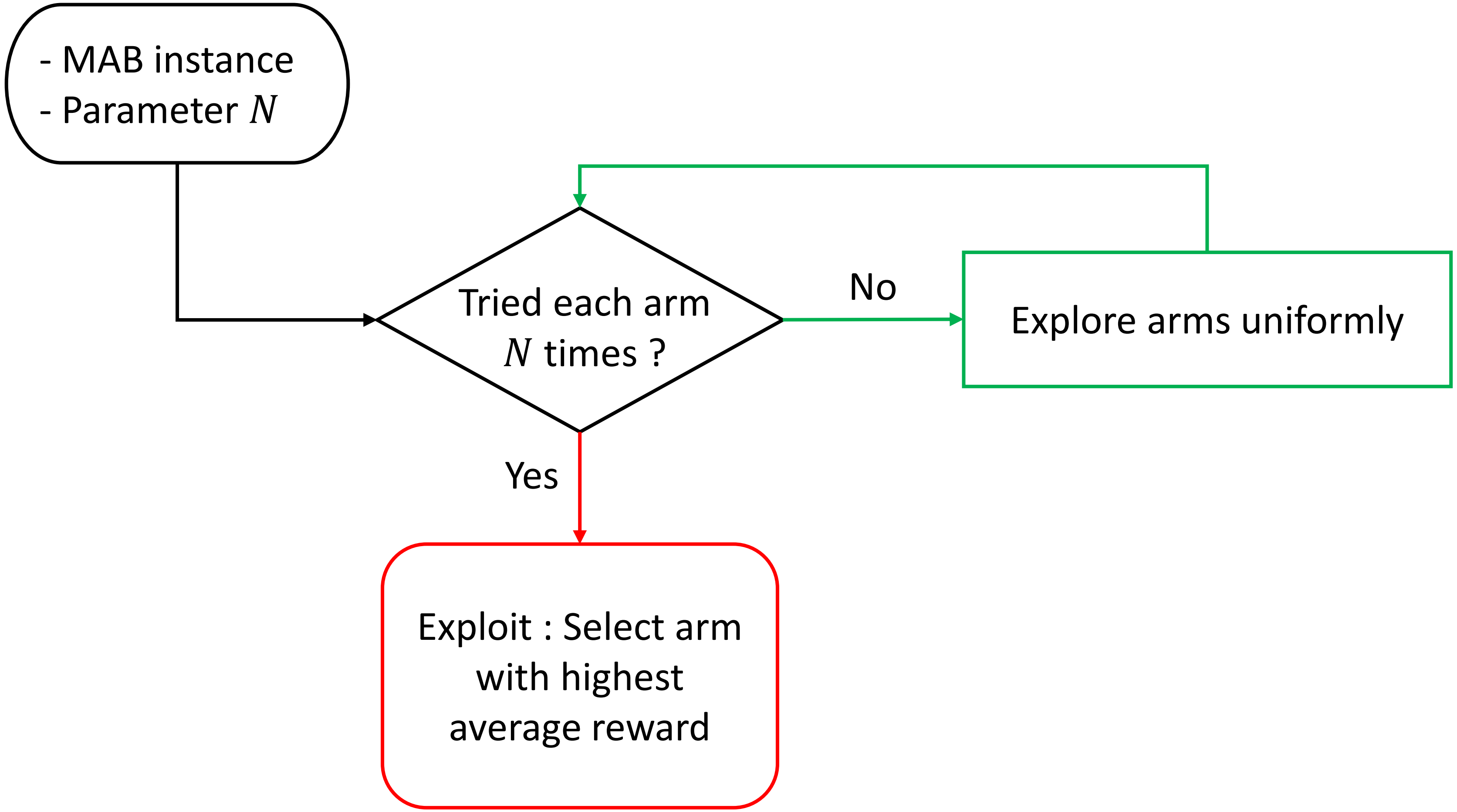
Basic protocol :

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

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

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

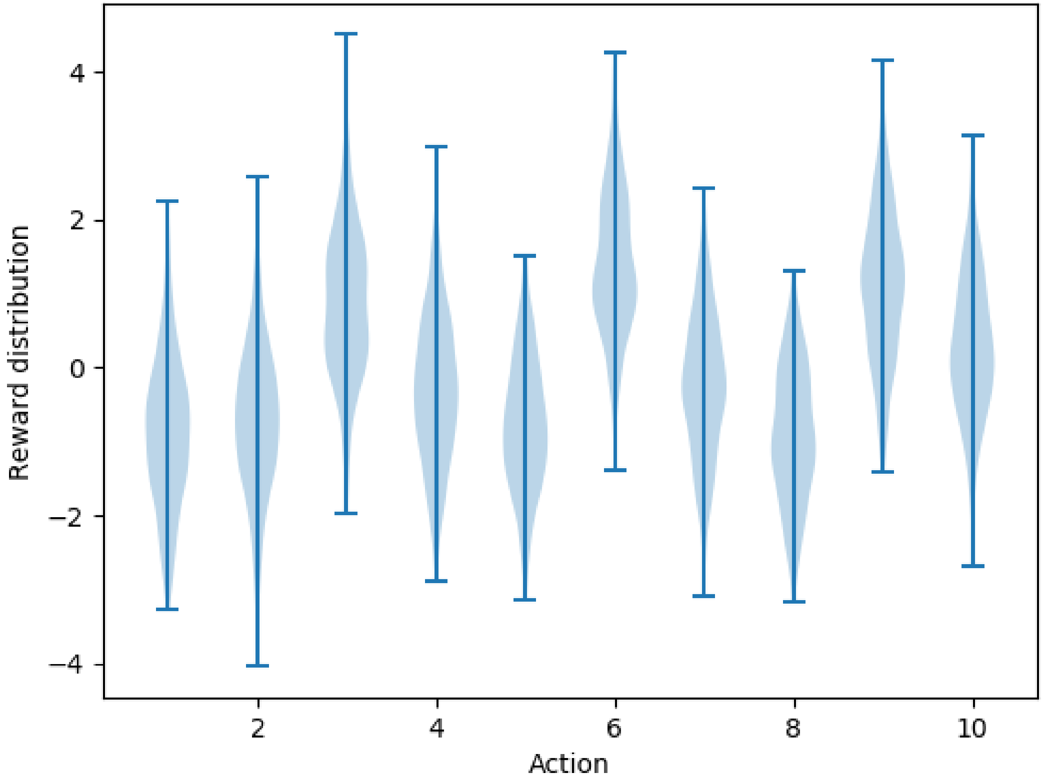
minimize with $N = T^{\frac{2}{3}}(\log T)^{\frac{1}{3}}$

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

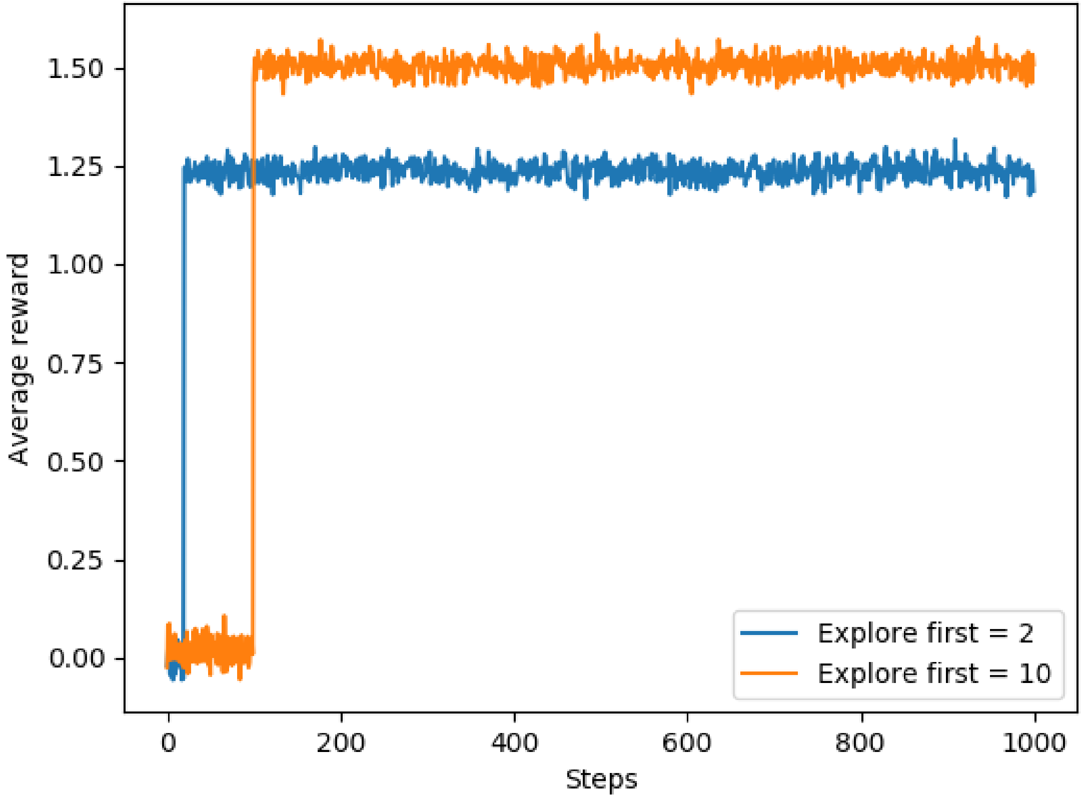
Regret Analysis

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

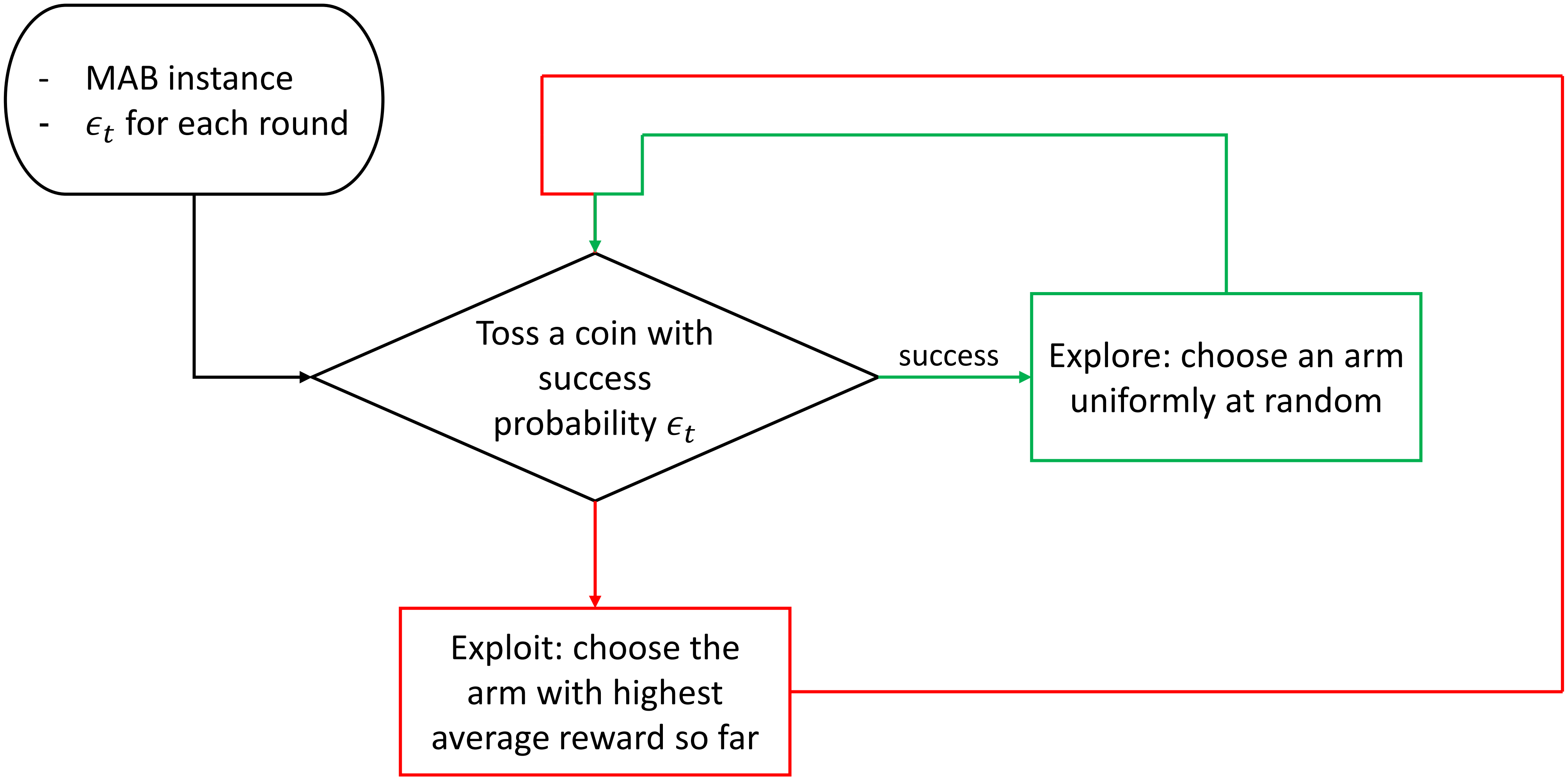
10-armed Testbed



Benchmark: 10-armed Testbed



ϵ-greedy

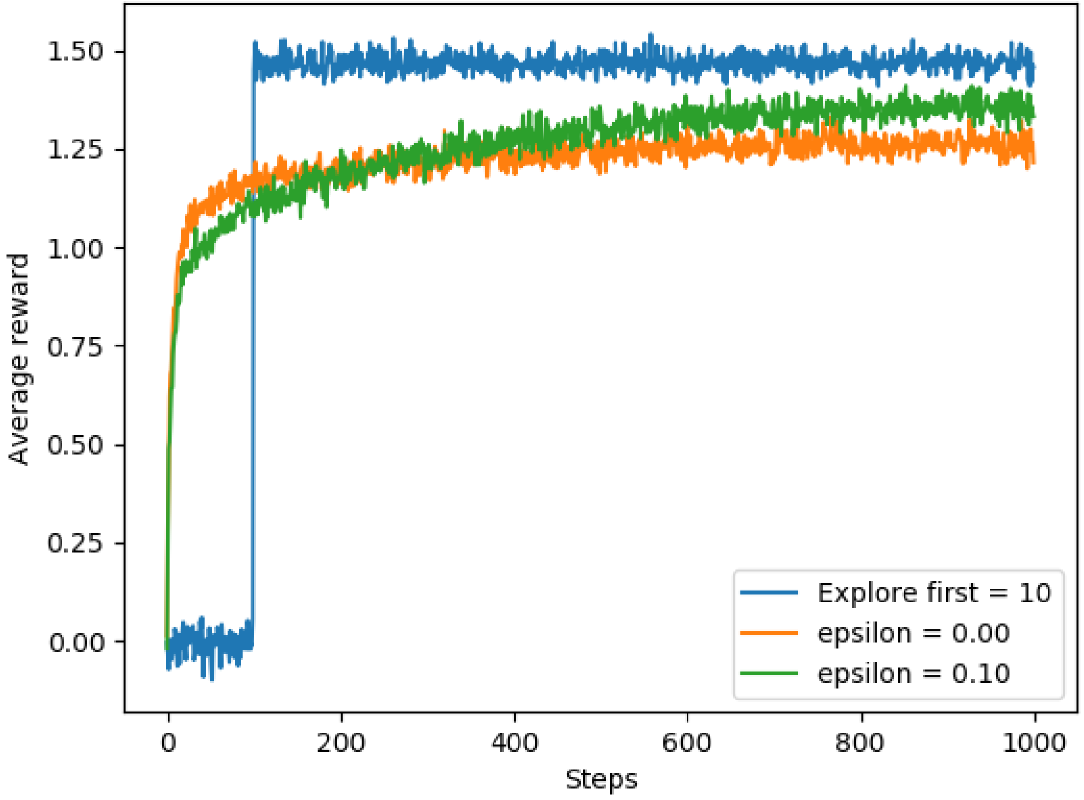


Regret Analysis

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

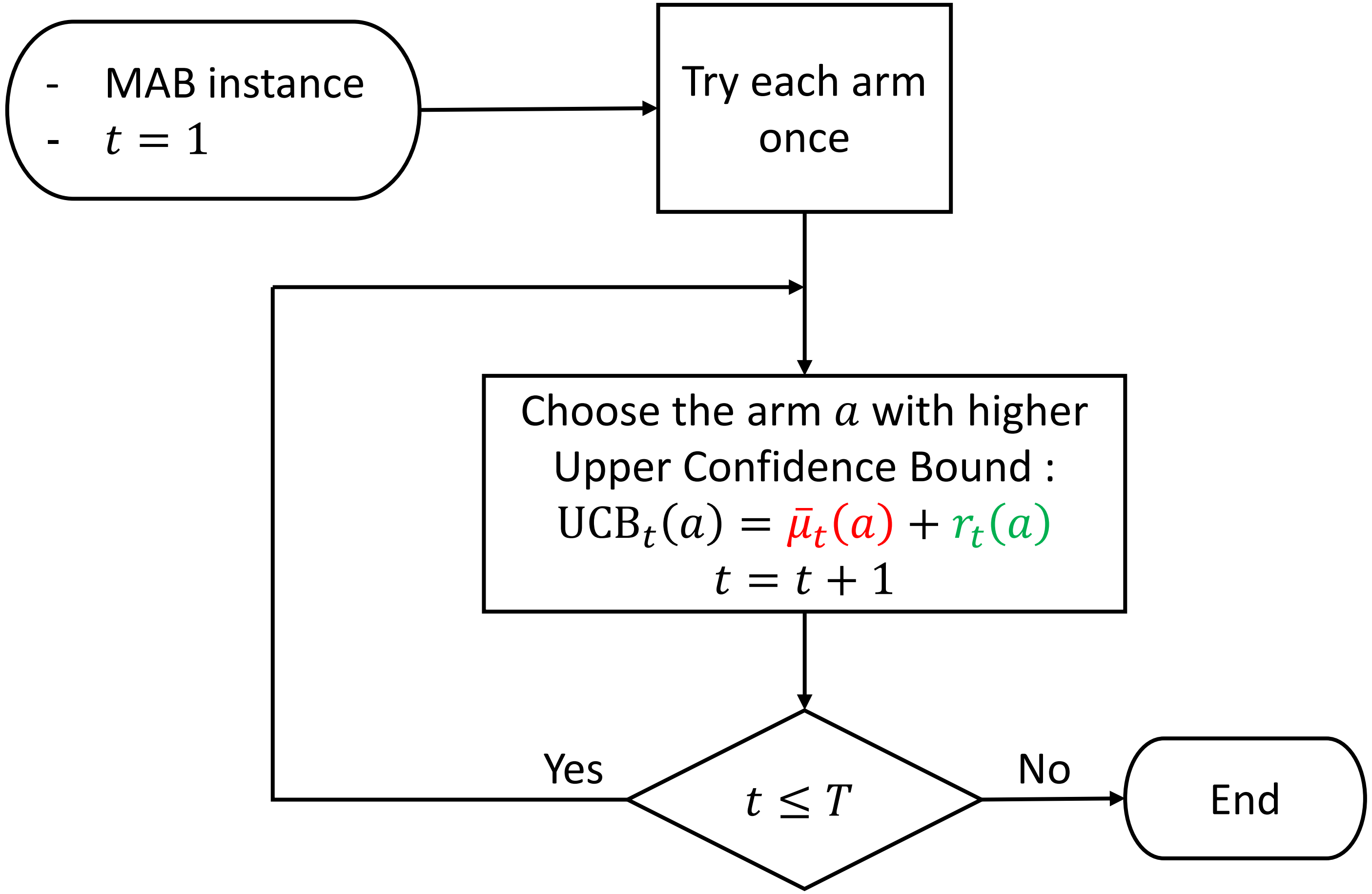
$$\mathbb{E}[R(t)] \leq t^{\frac{2}{3}} \times O(K \log t)^{\frac{1}{3}}$$

Benchmark



1.50
1.25
1.00
0.75
0.50
0.25
0.00
Average reward
0
200
400
600
800
1000
Steps
Explore first = 10
epsilon = 0.00
epsilon = 0.10

Upper Confidence Bound exploration



Upper Confidence Bound exploration

UCB1:

Try each arm once:

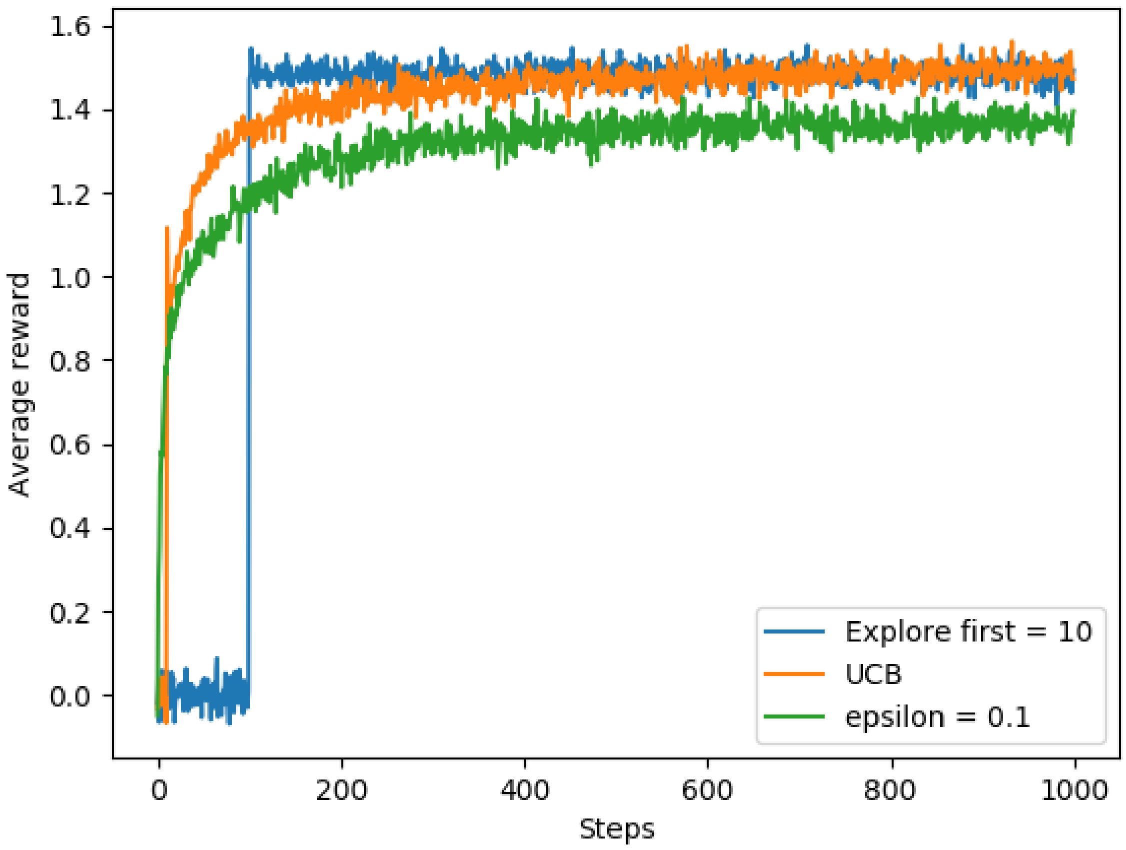
In each round t, pick $\underset{a \in A}{\operatorname{argmax}} UCB_t(a)$, where $UCB_t(a) = \bar{\mu}_t(a) + r_t(a)$

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

Regret Analysis

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

Benchmark


1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Average reward
0
200
400
600
800
1000
Steps
Explore first = 10
UCB
epsilon = 0.1

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

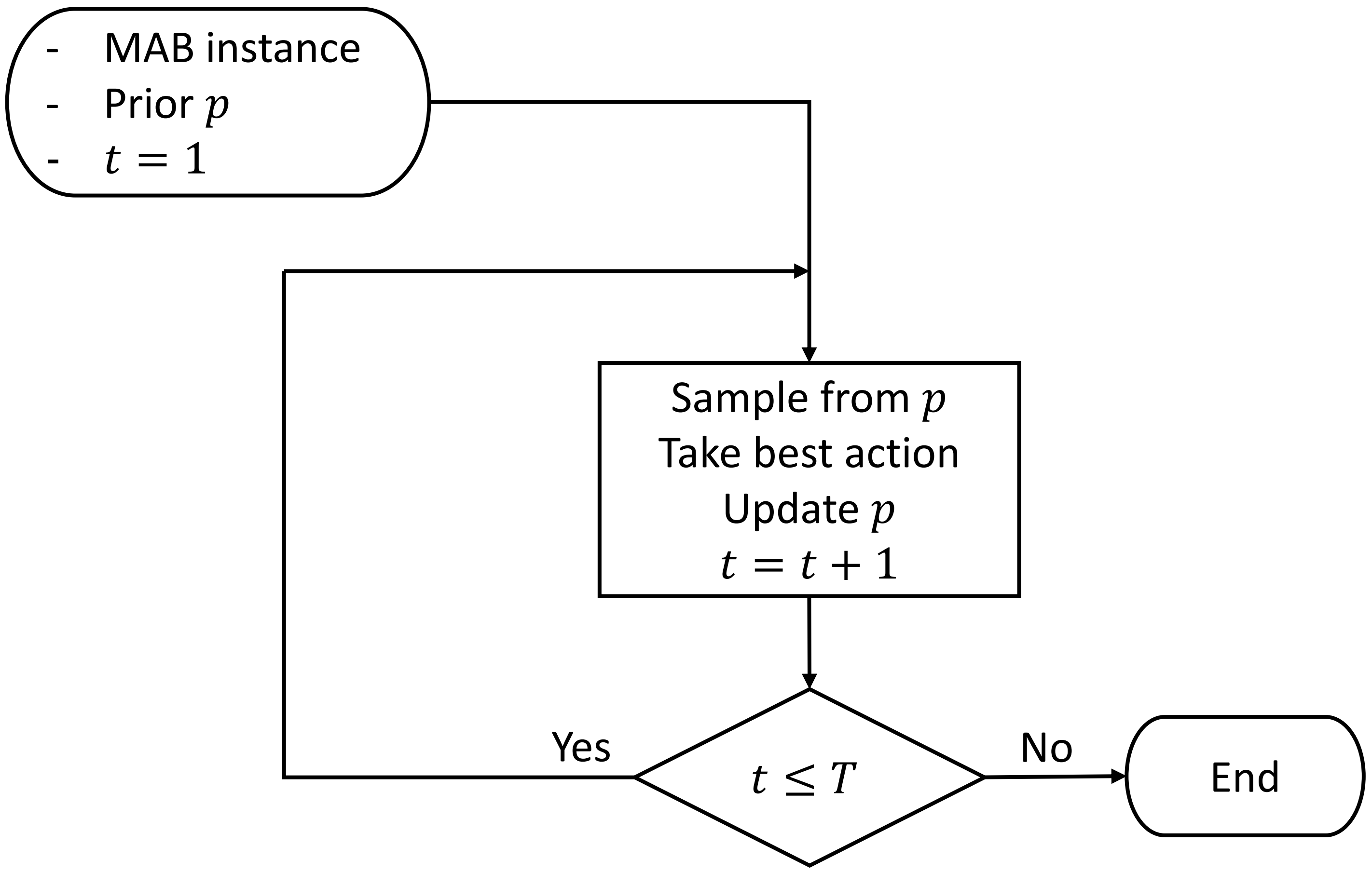
Bayesian assumption : $I \sim \mathbb{P}$

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

Bayesian regret:

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

Thompson Sampling



Terminology

t-history:
$H_t = \big((a_1, r_1), \ldots, (a_t, r_t)\big) \in (A \times \mathbb{R})^t$

feasible t-history:
$H = \big((a'_1, r'_1), \ldots, (a'_t, r'_t)\big) \in (A \times \mathbb{R})^t$
with $\Pr[H_t = H] > 0$

Thompson Sampling

For each round t

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

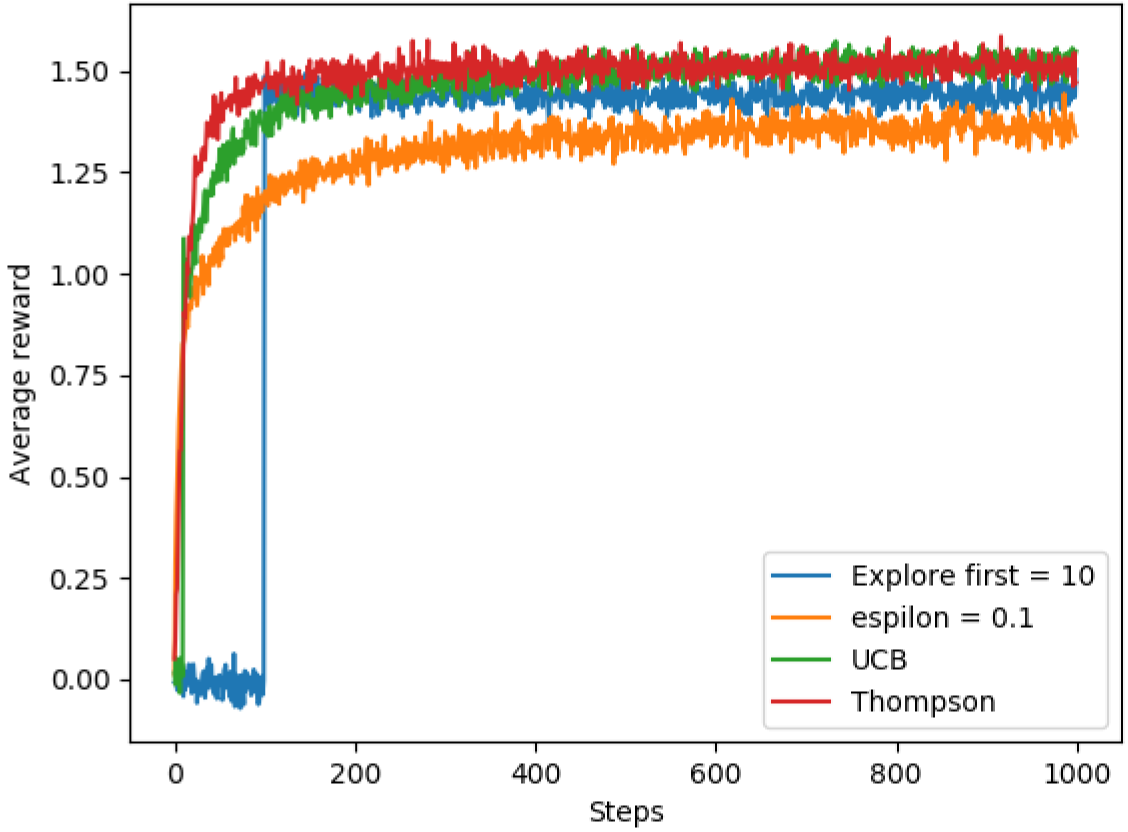
Draw arm a_t independently from distribution $p_t(\cdot\,|H)$, where

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

Bayesian regret analysis

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

Benchmark



1.50
1.25
1.00
0.75
0.50
0.25
0.00
Average reward
0
200
400
600
800
1000
Steps
Explore first = 10
espilon = 0.1
UCB
Thompson

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:

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

Fixed discretization

Discretization:
Finite set of arms $S \subset X$

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

Discretizatin error:
$\mathrm{DE}(S) = \mu^*(X) - \mu^*(S)$

Regret Analysis

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

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

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

Regret Analysis

Fixed uniform discretization :

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

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

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

Lipschitz MAB

Recall CAB:

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

Now:

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

Regret Analysis

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

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

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

Regret Analysis

Recall:

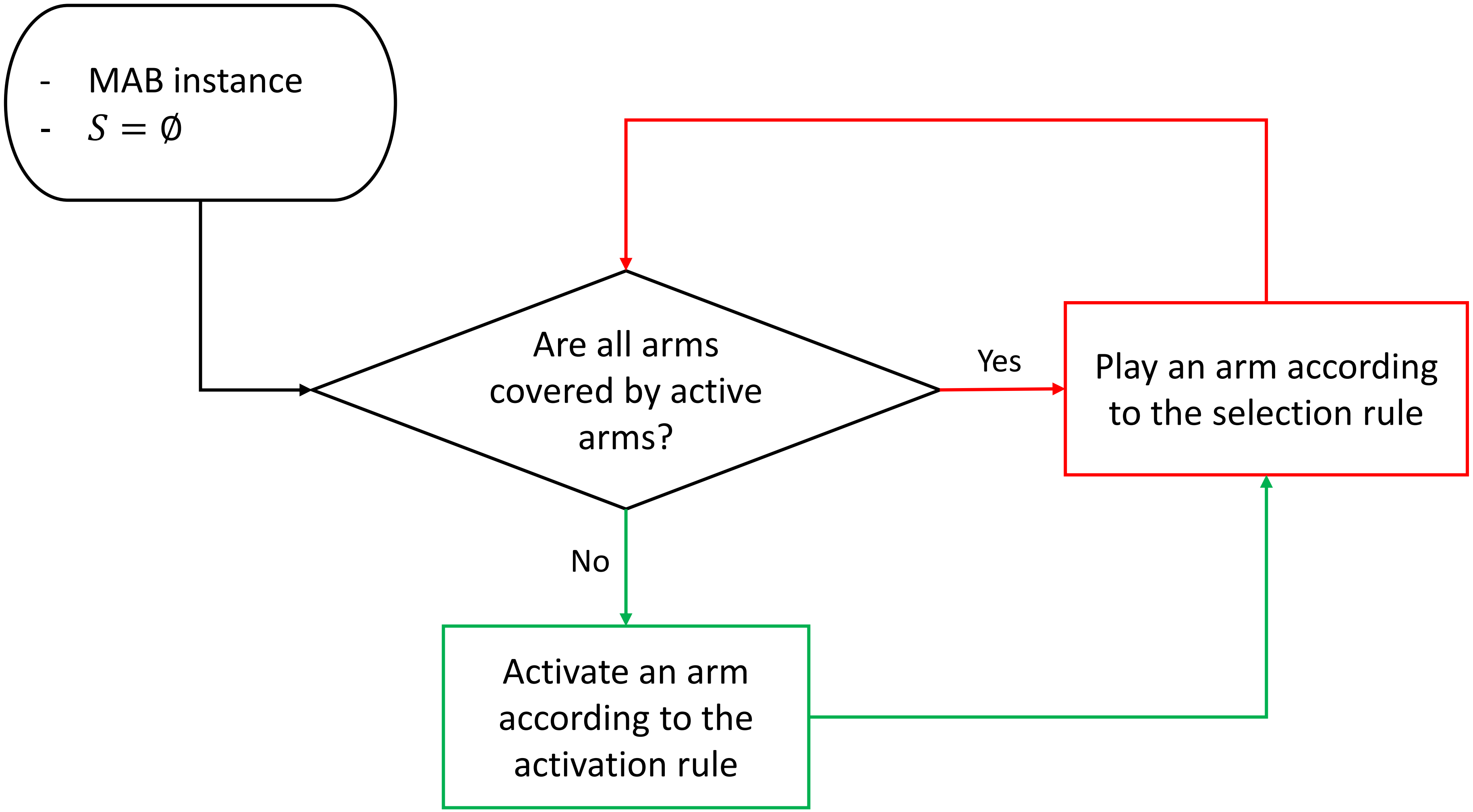
$$\mathbb{E}[R(T)] \leq O\left(\sqrt{|S| T \log T}\right) + T \cdot \mathrm{DE}(S)$$

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

Adaptive discretization

$$\mathrm{DE}(S) \leq \mathcal{D}(S, x^*) \coloneqq \min_{x \in S} \mathcal{D}(x, x^*)$$

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 $\mathrm{index}_t(x)$

Activation rule

Selection rule

Zooming Algorithm: Notations

Confidence radius:

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

Confidence Ball:

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

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