Efficient Algorithms for Infinite-Armed Bandit

Arghya Roy Chaudhuri
under the guidance of
Prof. Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-
Round 3	1	-	-	-	-	-

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-
Round 3	1	-	-	-	-	-
Round 4	1	-	-	-	-	-

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-
Round 3	1	-	-	-	-	-
Round 4	1	-	-	-	-	-
Round 5	0	-	-	-	-	-

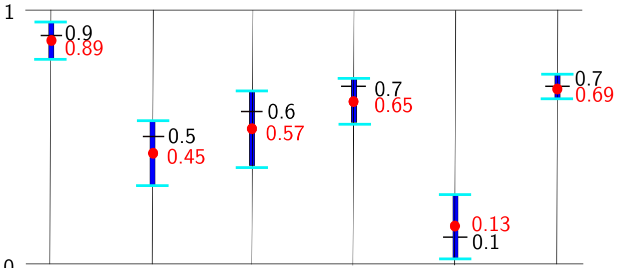
What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-
Round 3	1	-	-	-	-	-
Round 4	1	-	-	-	-	-
Round 5	0	-	-	-	-	-
Round 6	-	-	-	1	-	-

What is a Multi Armed Bandit ?

Machines:						
Mean Reward	0.9	0.5	0.6	0.7	0.1	0.7
Round 1	1	1	0	1	0	0
Round 2	-	0	-	-	-	-
Round 3	1	-	-	-	-	-
Round 4	1	-	-	-	-	-
Round 5	0	-	-	-	-	-
Round 6	-	-	-	1	-	-

Objective: Output the arm with the highest expected reward with high probability, while incurring a **minimal** number of samples

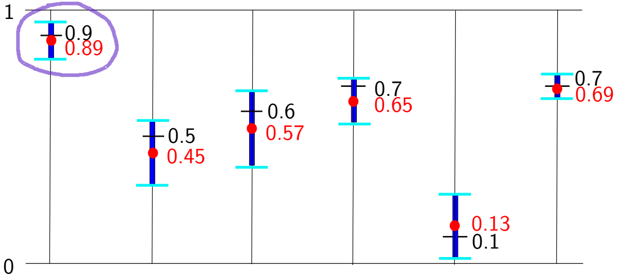


$$\underbrace{\hat{p} - \sqrt{\frac{2}{u}\ln\left(\frac{1}{\delta}\right)}}_{\text{Lower Confidence Bound(LCB)}} \leq p \leq \underbrace{\hat{p} + \sqrt{\frac{2}{u}\ln\left(\frac{1}{\delta}\right)}}_{\text{Upper Confidence Bound(UCB)}} \quad \textbf{w.p } 1-\delta$$

Approach:

- Track confidence bounds for each arm

Key Principle: Confidence Bounds



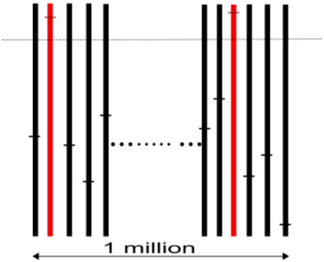
$$\underbrace{\hat{p} - \sqrt{\frac{2}{u} \ln\left(\frac{1}{\delta}\right)}}_{\text{Lower Confidence Bound(LCB)}} \quad \leq \quad p \leq \quad \underbrace{\hat{p} + \sqrt{\frac{2}{u} \ln\left(\frac{1}{\delta}\right)}}_{\text{Upper Confidence Bound(UCB)}} \quad \textbf{w.p} \;\; 1 - \delta$$

Approach:

- Track confidence bounds for each arm
- Return an arm whose LCB exceeds UCB of all the other arms

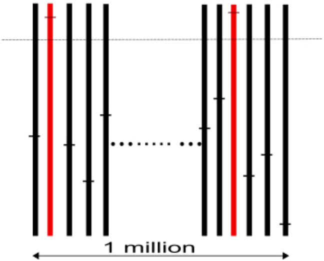
Our Problem

What if the number of arms is too large?



Our Problem

What if the number of arms is too large?



Problem Definition: Find an arm from an infinite set of arms whose expected reward is greater than $(1-\rho)^{\text{th}}$-quantile (for $0 < \rho < 1$) of distribution of rewards over arms.

Key to our Approach

Consider a biased coin with $P(\textbf{HEAD}) = \mathbf{0.1}$ and $P(TAIL) = 0.9$

Number of tosses	P(**no** Head)
1	0.9

Key to our Approach

Consider a biased coin with $P(\textbf{HEAD}) = \mathbf{0.1}$ and $P(TAIL) = 0.9$

Number of tosses	P(**no** Head)
1	0.9
10	0.348

Key to our Approach

Consider a biased coin with $P(\mathbf{HEAD}) = \mathbf{0.1}$ and $P(TAIL) = 0.9$

Number of tosses	P(**no** Head)
1	0.9
10	0.348
20	0.122

Key to our Approach

Consider a biased coin with $P(\mathbf{HEAD}) = \mathbf{0.1}$ and $P(TAIL) = 0.9$

Number of tosses	P(**no** Head)
1	0.9
10	0.348
20	0.122
50	0.005

Applications:

- Large/continuous action spaces with discontinuous rewards