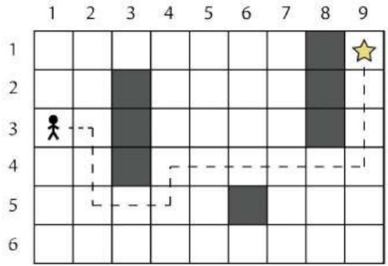
An introduction to Multi-armed bandit problem

Alireza Kavoosi
School of Industrial Engineering, University of Tehran

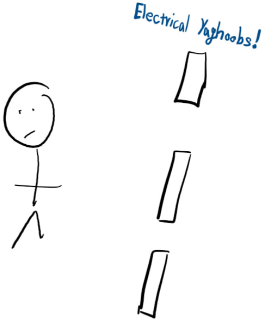
Outline

MAB or MDP?

MDP



MAB



Stochastic Bandits: Setup

- K arms, each arm k yields i.i.d. rewards $\{X_{k,t}\}$ with mean μ_k.
- Goal: Find the arm with the best expected reward $\mu^* = \max_k \mu_k$.
- A policy π selects an arm at each time t based on past observations.

Why is optimal exploration essential?

Why do we not want to stop exploring?

Regret

- Regret after n rounds:

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_{\pi_t,t}\right] = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}[T_k(n)].$$

- $\Delta_k = \mu^* - \mu_k$ is the gap for arm k.

Why are we working with "regret" instead of "reward"?

What does high regret tell you about your exploration-exploitation balance?

Warm Up: Full Information ($K = 2$)

- Observe outcomes $\{X_{1,t}, X_{2,t}\}$ after pulling any arm.
- Empirical mean:

$$\bar{X}_{k,t} = \frac{1}{t}\sum_{s=1}^{t} X_{k,s}.$$

- Choose the arm with highest $\bar{X}_{k,t}$.

How much information is enough?

SubGaussian Assumption

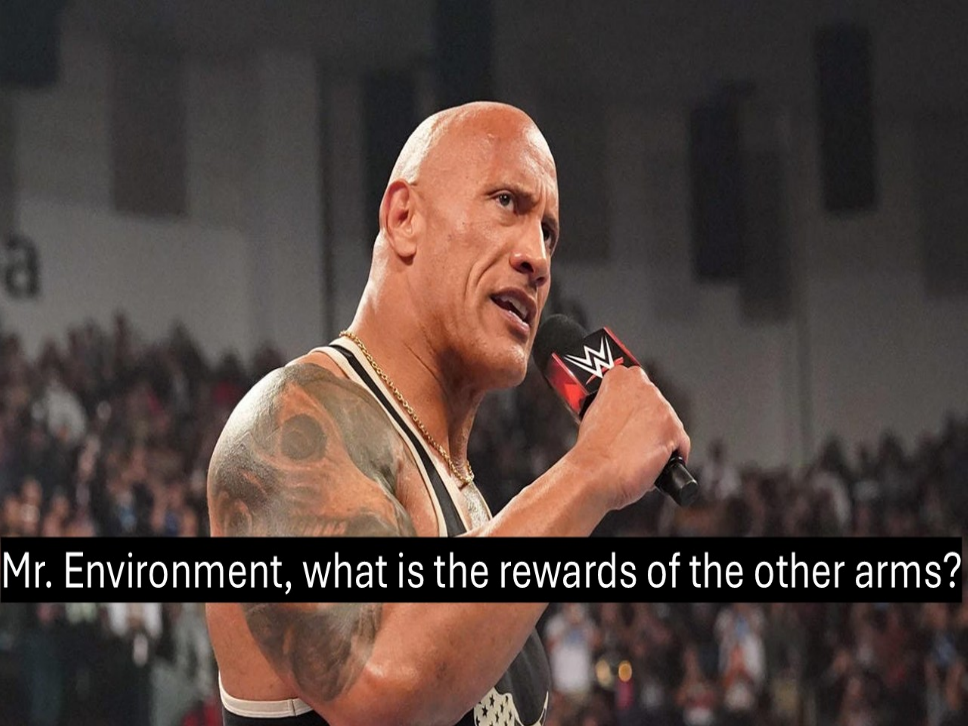
- Assume $X_{k,t}$ are subGaussian with proxy variance σ^2:

$$\mathbb{E}\left[e^{u(X_{k,t}-\mu_k)}\right] \leq e^{\frac{u^2\sigma^2}{2}}.$$

- Chernoff bounds yield regret:

$$R_n \leq \Delta + \frac{4\sigma^2}{\Delta}.$$

How does uncertainty in estimates affect decisions when means are close?


WWE
Mr. Environment, what is the rewards of the other arms?


$4
$6
$3
$8
$7
$666
Their rewards was ...
$$$$$$$$$$$$$$$


I BRING IT
IT DOESN’T MATTER HOW THEY’RE PAYING!

Upper Confidence Bound (UCB)

- Highest empirical mean can mislead if an arm is under-sampled.
- Boost empirical mean with an exploration bonus.

When might a high empirical mean be misleading?

UCB Strategy

- After $T_k(t)$ pulls:

$$\hat{\mu}_{k,t} = \frac{1}{T_k(t)} \sum_{s:\pi_s=k} X_{k,s}.$$

- Select arm:

$$\pi_t \in \arg\max_k \left\{ \hat{\mu}_{k,t} + 2\sqrt{\frac{\log t}{T_k(t)}} \right\}.$$

Why does the exploration bonus shrink with more pulls?

Algorithm: UCB

Algorithm 1 Upper Confidence Bound (UCB)

Input: K, n
for $t = 1$ to K **do**
    Pull each arm once.
end for
for $t = K + 1$ to n **do**
    Choose arm maximizing $\hat{\mu}_{k,t} + 2\sqrt{\frac{\log t}{T_k(t)}}$.
end for

UCB Regret Analysis

- UCB achieves:

$$R_n \leq \sum_{\Delta_k > 0} \frac{8 \log n}{\Delta_k} + \left(1 + \frac{\pi^2}{3}\right) \Delta_k.$$

- Trade-off: Small gaps Δ_k make distinguishing arms harder.

What trade-off is captured in this regret bound?

What is our biggest concern in the UCB algorithm?

Bounded Regret Policy (BRP)

- Can regret be bounded independent of n?
- Assume gap Δ is known.
- Set:

$$\mu_1 = \frac{\Delta}{2}, \quad \mu_2 = -\frac{\Delta}{2}.$$

Under what conditions can regret remain bounded?

Algorithm: BRP for $K = 2$

Algorithm 2 Bounded Regret Policy (BRP)

```
Pull each arm once.
for t = 3 to n do
    if max_k μ̂_{k,t} > 0 then
        Pull arm with highest μ̂_{k,t}.
    else
        Alternate arms.
    end if
end for
```

BRP Regret Bound

- Regret is bounded:
$$R_n \leq \Delta + \frac{16}{\Delta}.$$
- Key: Use sign of empirical mean to decide early.

How does knowing the gap simplify the learning process?

BRP Error Sources

- Two error types:
  1. Suboptimal arm appears optimal.
  2. Optimal arm appears suboptimal.
- Analyzed via union bound and Chernoff bounds.

What are the main sources of error in estimation?

Conclusion

(What key insights guide exploration vs. exploitation?)

- Reviewed multi-armed bandits and regret.
- Two methods discussed:
  1. UCB: Logarithmic regret through exploration bonuses.
  2. BRP: Constant regret in a controlled, two-arm setting (assuming known gap).

Think

How do these strategies guide real-world decision-making?

References

- **Lecture Notes:** MIT 18.657 Mathematics of Machine Learning, Fall 2015.
- Lattimore, T. (2015) *Optimally Confident UCB: Improved regret for finite-armed bandits.* `http://arxiv.org/abs/1507.07880`
- MIT OpenCourseWare: `http://ocw.mit.edu/`

Thank you for your attention!