CS155/254: Probabilistic Methods in Computer Science

The Multi-Armed Bandit

Gambling in a Rigged Casino

- A collection of slot machines giving random rewards.
- Rewards' distributions of different machines may be different.
- Rewards' distributions unknown to the player.

How to maximize total reward in a sequence of T actions (arm pulls)?

A good strategy must balance the tradeoff between:

1. **Exploit:** play arms that seem best based on current information
2. **Explore:** try other arms to get more information on possibly better arms

The Multi-Armed Bandit Problem

- Multi-armed bandit problems were studied for over 70 years in economics, operation research and (more recently) is computer science, as an abstraction of Reinforcement learning (RL).
- Applications:
  - Clinical trials – find the best of several experimental treatments while minimizing damage to patients.
  - Adaptive routing - minimize delays by exploring alternative paths.
  - Finance/Investment: optimal asset allocation.
  - Production: scheduling, resource allocation.
  - Economics: pricing a product.
  - Web: content matching, efficient crawling.
  - Robotics

Stochastic Multi-Armed Bandit

- Set of k arms (actions)
- A set of k **unknown** expectations $\mu_1, \ldots, \mu_k$.
- The payoff of arm i is a random variable $X_i \in [0, 1]$, with expectation $E[X_{i,}] = \mu_i$.
- Successive payoffs are independent events.
- Let $i^* = \arg\max_{j \in [k]} \mu_i$ be the optimal arm,
- Let i_t be the arm pulled at step t.
- Given a sequence of T actions $S(T) = i_1, \ldots, i_T$, the expected "regret" of this sequence is

$$E[S(T)] = \sum_{t=1}^{T} (\mu_{i^*} - \mu_{i_t})$$

Our goal is to minimize $E[S(T)]$.

First Attempt: Explore and Commit

Algorithm EaC:

1. Pull each arm m times.
2. Let M_i be the empirical mean of arm i in the m activations.
3. After the first km steps always activate arm $j = \arg\max_{i \in [k]} M_i$.

Let $\Delta_i = \mu_{i^*} - \mu_i$ be the expected loss of activating arm i.

Theorem

The expected regret of Algorithm EaC is

$$= \sum_{i=1}^{k} m\Delta_i + (T - km) \sum_{j=1}^{k} \Delta_j Pr\left(j = \arg\max_{i \in [k]} M_i\right)$$

For a fixed m, $Pr(M_i > M_{i^*}) > 0$, thus $\sum_{j=1}^{k} \Delta_j Pr\left(j = \arg\max_{i \in [k]} M_i\right) > 0$.

First Attempt: Explore and Commit

Algorithm EaC:

1. Pull each arm m times.
2. Let M_i be the empirical mean of arm i in the m activations.
3. After the first km steps always activate arm $j = \arg\max_{i \in [k]} M_i$.

Let $\Delta_i = \mu_{i^*} - \mu_i$ be the expected loss of activating arm i.

Theorem

The expected regret of Algorithm EaC is

$$= \sum_{j=1}^{k} m\Delta_j + (T - km) \sum_{j=1}^{k} \Delta_j Pr\left(j = \arg\max_{i \in [k]} M_i\right) = B + (T - km)C$$

where $B > 0$ and $C > 0$ that are constants independent of T.

The expected regret is linear in T, since the algorithm stops learning after a fixed (km) number of steps.

Second Attempt: β-Greedy Algorithm

Algorithm Greedy(β):

1. Set $M_i = 0$ for all arms.
2. Repeat
   1. With probability β choose j uniformly at random, else let $j = \arg\max_{i \in [k]} M_i$.
   2. Active arm j and update M_j.

Theorem

The expected regret of the β-Greedy Algorithm is

$$\geq \frac{\beta T}{k} \sum_{i=1}^{k} \Delta_i$$

The expected regret is linear in T because the algorithm continues to explore after finding the optimal arm.

The Explore and Commit algorithm stops exploring too early.

The β-Greedy algorithm continues to explore even after finding an optimal strategy.

Both algorithm don't adapt their moves to the accuracy of their estimates.

Third Attempt: Upper Confidence Bound (UCB) Algorithm

Algorithm UCB(α):

1. For all $i \in [k]$ activate arm i once, update M_i and set $N_i = 1$.
2. For $t = k+1, \ldots, T$ do:
   1. Activate arm $j = \arg\max_{i \in [k]} \left(M_i + \sqrt{\frac{\alpha \log t}{2N_i}} \right)$.
   2. Update M_j, and set $N_j = N_j + 1$

Theorem

The expected regret of UCB(α), $\alpha > 1$ is bounded by

$$\sum_{i \in [k],\ \Delta_i > 0} \left(\frac{2\Delta_i}{\alpha - 1} + \frac{2\alpha \log T}{\Delta_i} \right)$$

The regret grows logarithmically in T. This growth is asymptotically optimal for this setting.

Confidence Interval

Definition

An (ϵ, δ)-**confidence interval** for a constant V is a random variable ("estimator") $\tilde{V}$ such that

$$Pr(\tilde{V} - \epsilon \leq V \leq \tilde{V} + \epsilon) \geq 1 - \delta.$$

Let $M_i(m)$ be the average reward of arm i in m activations.
We estimate μ_i with the estimator $M_i(m)$,

Applying Hoeffding's inequality, we have

$$Pr(|M_i(m) - \mu_i| \geq \epsilon) \leq 2e^{-2m\epsilon^2}$$

For any $\delta < 1$,

$$Pr\left(\mu_i - \sqrt{\frac{-\log \delta}{2m}} \leq M_i(m) \leq \mu_i + \sqrt{\frac{-\log \delta}{2m}}\right) \geq 1 - \delta.$$

Note that the width of the interval decreases with the number of trials m.

Algorithm UCB(α):

1. For all $i \in [k]$ activate arm i once, update M_i and set $N_i = 1$.
2. For $t = k+1, \ldots, T$ do:
   1. Activate arm $j = \arg\max_{i \in [k]} \left(M_i + \sqrt{\frac{\alpha \log t}{2N_i}} \right)$.
   2. Update M_j, and set $N_j = N_j + 1$

$$Pr\left(\mu_i - \sqrt{\frac{-\log \delta}{2m}} \leq M_i(m) \leq \mu_i + \sqrt{\frac{-\log \delta}{2m}} \right) \geq 1 - \delta.$$

UCB maintains confidence intervals for the estimates M_i's of the constants μ_i's.
The estimate is updated (in M_i and N_i) when an arm is activated, reducing the size of the interval.
The value of δ is chosen for a union bound over the steps.

Consider arm i, with $\Delta_i > 0$.
Let $M_i(s)$ and $N_i(s)$ be the values of the variables M_i and N_i at iteration s.
If arm i was activated at iteration s, then

$$M_i(s) + \sqrt{\frac{\alpha \log s}{2N_i(s)}} \geq M_{i^*}(s) + \sqrt{\frac{\alpha \log s}{2N_{i^*}(s)}}$$

But,

$$Pr\left(\mu_i - \sqrt{\frac{\alpha \log s}{2N_i(s)}} \leq M_i(s) \leq \mu_i + \sqrt{\frac{\alpha \log s}{2N_i(s)}}\right) \geq 1 - s^{-\alpha}.$$

If $N_i(s) > \frac{2\alpha \log s}{\Delta_i^2}$, then with probability $1 - 2s^{-\alpha}$,

$$M_i(s) + \sqrt{\frac{\alpha \log s}{2N_i(s)}} \leq \mu_i + 2\sqrt{\frac{\alpha \log s}{2N_i(s)}} < \mu_i + \Delta_i \leq \mu_{i^*} \leq M_{i^*}(s) + \sqrt{\frac{\alpha \log s}{2N_{i^*}(s)}}$$

If $N_i(s) > \frac{2\alpha \log s}{\Delta_i^2}$, i is activated with probability $\leq 2s^{-\alpha}$.

Consider arm i, with $\Delta_i > 0$.

Lemma

Let $N_i(t)$ be the number of times arm i was activated in the first t iterations. Then

$$E[N_i(t)] \leq \frac{2\alpha \log t}{\Delta_i^2} + \frac{2}{\alpha - 1}.$$

Proof: If $N_i(s) \geq \frac{2\alpha \log s}{\Delta_i^2}$, i is activated with probability $2s^{-\alpha}$.

$$E[N_i(t)] \leq \frac{2\alpha \log t}{\Delta_i^2} + \sum_{s \geq \frac{2\alpha \log t}{\Delta_i^2}} 2s^{-\alpha} \leq \frac{2\alpha \log t}{\Delta_i^2} + \frac{2}{\alpha - 1}$$

$$E[N_i(t)] \leq \frac{2\alpha \log t}{\Delta_i^2} + \sum_{s \geq \frac{2\alpha \log t}{\Delta_i^2}} 2s^{-\alpha} \leq \frac{2\alpha \log t}{\Delta_i^2} + \frac{2}{\alpha - 1}$$

The expected regret in T steps bounded by

$$\sum_{i \in [k],\ i \neq i^*} E[N_i(T)]\Delta_i \leq \sum_{i \in [k],\ i \neq i^*} \left(\frac{2\alpha \log T}{\Delta_i} + \frac{2\Delta_i}{\alpha - 1} \right)$$