CS 4/5789 28 Mar 2022 Prof Sarah Dean

Lecture 18: Multi-Armed Bandits & Confidence Bounds

1) Explore-then-Commit

Alg 3: Explore-then-commit:

For $t = 1, \dots, N \cdot K$ } pull each arm N times
  $a_t = t \bmod K$

$\hat{\mu}_a = \frac{1}{N} \sum_{i=1}^{N} r_{Ki}$ } compute average reward

For $t = N \cdot K + 1, \dots, T$
  $a_t = \arg\max_a \hat{\mu}_a = \hat{a}^*$

This algorithm balances exploration & exploitation.

How to set N?

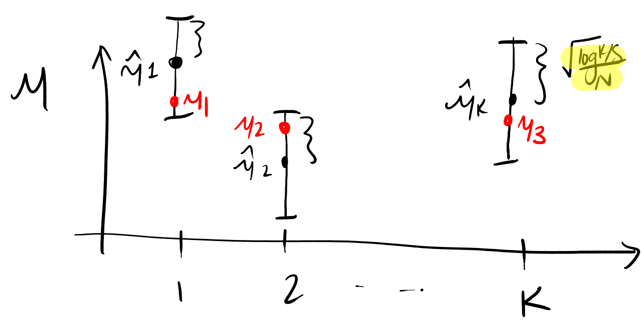
Let's do some analysis.

The regret decomposes:

$$R(T) = \sum_{t=1}^{T} \mu^* - \mu_{a_t} = \underbrace{\sum_{t=1}^{NK} \mu^* - \mu_{a_t}}_{R_1} + \underbrace{\sum_{t=NK+1}^{T} \mu^* - \mu_{\hat{a}^*}}_{R_2}$$

To bound R_2, consider the difference between $\hat{\mu}_a$ and μ_a.

We suppose rewards are bounded $r_t \in [0, 1]$.



Lemma (Explore): After exploration phase, for all arms $a = 1, \dots, K$,

$$|\hat{\mu}_a - \mu_a| \lesssim \sqrt{\frac{\log(K/\delta)}{N}} \quad \text{with probability } 1-\delta.$$

Proof: Hoeffding & union bound $P(A \cap B) \leq P(A) + P(B)$.

Lemma (Hoeffdings): Suppose $r_i \in [0,1]$ and $\mathbb{E}[r_i] = \mu$. Then for $r_1, \dots, r_N$ iid, with probability $1-\delta$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} r_i - \mu \right| \lesssim \sqrt{\frac{\log(1/\delta)}{N}}$$ (proof is out of scope.)

This gives us $1-\delta$ confidence intervals: $\mu_a \in \left[\hat{\mu}_a \pm c\sqrt{\frac{\log(K/\delta)}{N}} \right]$

We use confidence intervals to bound R_2.

$$R_2 = \sum_{t=NK+1}^{T} \mu^* - \mu_{\hat{a}^*} = (T-NK)(\mu^* - \mu_{\hat{a}^*})$$

$$\le (T-NK)\left[\underbrace{\hat{\mu}_{a^*} + \sqrt{\frac{\log(K/\delta)}{N}}}_{\text{upper confidence bound}} - \Big(\underbrace{\hat{\mu}_{\hat{a}^*} - \sqrt{\frac{\log(K/\delta)}{N}}}_{\text{lower confidence bound}} \Big) \right]$$

$$= (T-NK)\Big(\underbrace{\hat{\mu}_{a^*} - \hat{\mu}_{\hat{a}^*}}_{\le 0 \text{ by definition of } \hat{a}^*} + 2\sqrt{\frac{\log(K/\delta)}{N}} \Big)$$

Combining everything, we have

$$R(T) = R_1 + R_2 \le \underbrace{NK}_{\text{explore cost}} + \underbrace{2T\sqrt{\frac{\log K/\delta}{N}}}_{\text{exploit cost (if wrong)}} \quad \text{w.p. } 1-\delta$$

Minimizing this upper bound with respect to N,

$N = \left(\frac{T}{2K} \sqrt{\log(K/\delta)} \right)^{2/3}$ and w.p. $1-\delta$,

$R(T) \lesssim T^{2/3} K^{1/3} \log^{1/3}\left(\frac{K}{\delta}\right)$ for explore-then-commit

sublinear!

2) Upper Confidence Bound Algorithm

Idea: always pull the arm that has the highest upper confidence bound.

Follows principle of optimism in the face of uncertainty

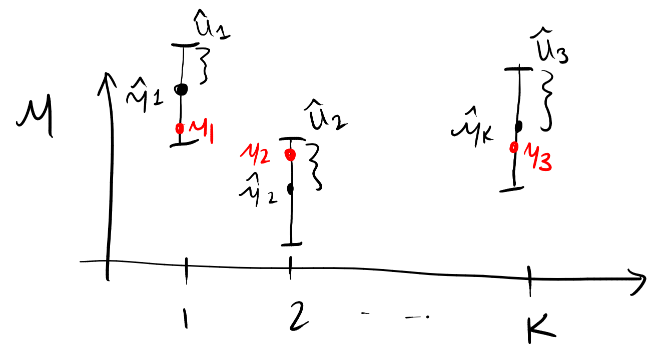
Alg 4: UCB

Initialize $\hat{\mu}_0^a$, N_0^a for $a=1,\dots,K$

For $t=1,2,\dots,T$:

$a_t = \arg\max_a \hat{u}_t^a \leftarrow = \hat{\mu}_t^a + \sqrt{\frac{\log(KT/\delta)}{N_t^a}}$

update $\hat{\mu}_{t+1}^{a_t}$ and $N_{t+1}^{a_t}$



The confidence intervals depend on # times an arm is pulled

$$N_t^a = \sum_{k=1}^{t} \mathbb{1}\{a_t = a\}$$

Also depend on the empirical mean

$$\hat{\mu}_t^a = \sum_{k=1}^{t} r_t \mathbb{1}\{a_t = a\} / N_t^a$$

The $1-\delta$ upper confidence bounds are

$$\hat{u}_t^a = \hat{\mu}_t^a + \sqrt{\frac{\log(KT/\delta)}{N_a^t}}$$

Q: why $\log(KT/\delta)$?

hint: recall union bound

This is like adding a synthetic reward bonus inversely proportional to the # times we visit a state.

3) UCB Analysis

The intuition for why UCB works is that we are in one of two cases each time we pull an arm:

case 1) a_t has a large confidence interval → explore
so high uncertainty

case 2) a_t has small confidence interval → exploit
so good arm

Regret at time t:

$$\mu^* - \mu_{a_t} \leq \hat{u}_t^{a^*} - \mu_{a_t}$$ (true mean within confidence interval for all arms)

$$\leq \hat{u}_t^{a_t} - \mu_{a_t}$$ ($a_t = \text{argmax } \hat{u}_t^a$)

$$= \hat{\mu}_t^{a_t} + \sqrt{\frac{\log(TK/\delta)}{N_t^{a_t}}} - \mu_{a_t}$$ (definition)

$$\leq 2\sqrt{\frac{\log(TK/\delta)}{N_t^{a_t}}}$$ (lower confidence interval)

Putting it all together,

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

$$\leq 2\sqrt{\log(TK/\delta)} \sum_{t=1}^{T} \sqrt{1/N_t^{a_t}}$$

Claim: $\sum_{t=1}^{T} \sqrt{1/N_t^{a_t}} \leq \sqrt{KT}$

$$\leq 2\sqrt{KT \log(KT/\delta)}$$

sublinear regret! $O(\sqrt{T})$ vs. $O(T^{2/3})$ explore-then-commit.

<u>Proof of claim</u> (optional)

$$\sum_{t=1}^{T} \sqrt{1/N_t^{a_t}} = \sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{1}\{a_t = a\} \sqrt{1/N_t^{a_t}}$$ (indicator $=1$ for only one term of the sum)

$$= \sum_{a=1}^{K} \left(\sum_{t=1}^{T} \mathbb{1}\{a_t = a\} \sqrt{1/N_t^{a_t}} \right)$$ (switching summation order)

$$= \sum_{a=1}^{K} \left(\sum_{t=1}^{N_T^a} \sqrt{1/t} \right)$$ (indicator $=1$ whenever $N_t^{a_t}$ increments)

$$\leq \sum_{a=1}^{K} \sqrt{N_T^a}$$ ($\sum_{i=1}^{N} 1/\sqrt{i} \leq \sqrt{N}$ summation rule)

Aside: $\sum_{a=1}^{K} N_T^a = T$ because we pull one arm per round.

$$\frac{1}{K} \sum_{a=1}^{K} \sqrt{N_T^a} \leq \sqrt{\frac{1}{K} \sum_{a=1}^{K} N_T^a} = \sqrt{T/K}$$

Jensen's

Therefore, $$\sum_{t=1}^{T} \sqrt{1/N_t^{a_t}} \leq \sum_{a=1}^{K} \sqrt{N_T^a} \leq \sqrt{KT}$$

$\square$