Exploration: Regret and Multi-Armed Bandits

Roberto Capobianco

SAPIENZA
UNIVERSITÀ DI ROMA

Adapted from Wen Sun's slides

Exploration: the Big Pain of RL

Exploration-Exploitation Trade-off: should we make the best decision given current information, or should we collect more information? In other words: should I sacrifice something now to get more in the future? (chicken-egg problem)

We need to carefully and systematically explore

Exploration: the Big Pain of RL

Exploration-Exploitation Trade-off: should we make the best decision given current information, or should we collect more information? In other words: should I sacrifice something now to get more in the future? (chicken-egg problem)

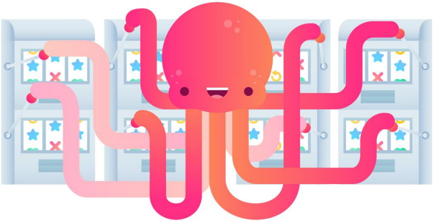
We need to carefully and systematically explore

e.g., go to my favourite restaurant vs try a new one

Multi-Armed Bandit

Let’s consider a simplified problem to analyze exploration:
Multi-Armed Bandits

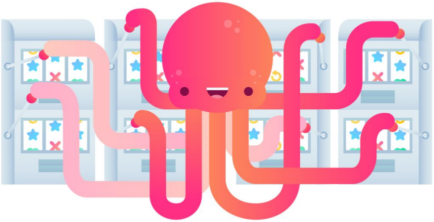
Multi-Armed Bandit



Let's consider a simplified problem to analyze exploration:
Multi-Armed Bandits

- We live in a world with no state, so decisions do not change based on some "context"
- K different arms (think of them as actions, or choices): a_1, ..., a_k
- A reward r is a scalar signal representing a feedback
  - We often assume it's in [0, 1]
- Each arm has unknown reward distribution ν_i with mean $\mu_i = \mathbb{E}_{r\sim\nu_i}[r]$
- Every time we pull an arm we observe an i.i.d. reward

Multi-Armed Bandit: Example



One domain of application of multi-armed bandits is online ads:

- Arms correspond to ads
- Each arm has a click-through-rate (0/1 reward based on click) that we aim to maximize

How do we decide which ad to propose next?

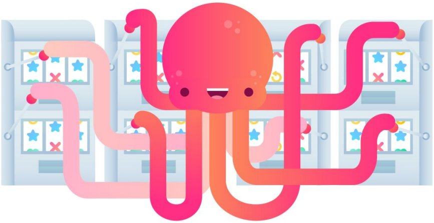
Multi-Armed Bandit: Interaction

The interactive process that we deal with in MAB is the following:

For $t = 0, \ldots, T-1$:

1. Pull an arm I_t in $\{1, \ldots, K\}$ based on historical information
2. Observe i.i.d. reward $r_i \sim \nu_i$ of arm I_t (we do not observe rewards of untried arms)

Multi-Armed Bandit: Interaction



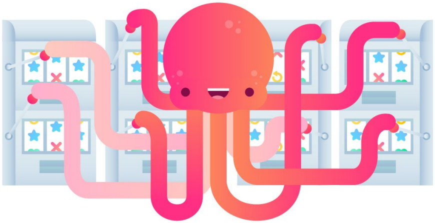
The interactive process that we deal with in MAB is the following:

For $t = 0, \ldots, T-1$:

1. Pull an arm I_t in $\{1, \ldots, K\}$ based on historical information
2. Observe i.i.d. reward $r_i \sim \nu_i$ of arm I_t (we do not observe rewards of untried arms)

But what are we trying to optimize exactly?

Multi-Armed Bandit: Interaction



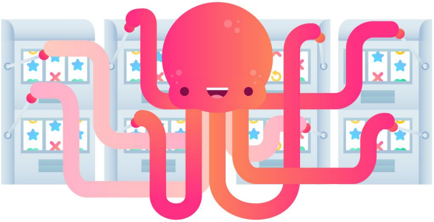
The interactive process that we deal with in MAB is the following:

For $t = 0, \ldots, T-1$:

1. Pull an arm I_t in $\{1, \ldots, K\}$ based on historical information
2. Observe i.i.d. reward $r_i \sim \nu_i$ of arm I_t (we do not observe rewards of untried arms)

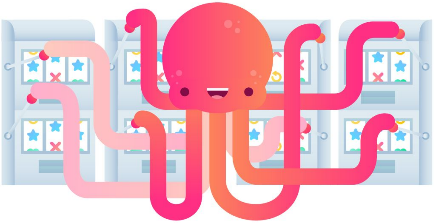
But what are we trying to optimize exactly? **REGRET!**

Regret



We want to minimize our **opportunity loss**, which is expressed in the form of the regret

Regret



We want to minimize our **opportunity loss**, which is expressed in the form of the regret

Assume we know what is the best arm to pull and its mean reward distribution μ^*

$$\mu^\star = \max_{i \in [K]} \mu_i$$

Regret

We want to minimize our **opportunity loss**, which is expressed in the form of the regret

The regret is the **total expected reward if we pull the best arm for T rounds** VS the **total expected reward of the arms we pulled over T rounds**

$$\text{Regret}_T = \boxed{T\mu^\star} - \boxed{\sum_{t=0}^{T-1} \mu_{I_t}} \qquad \mu^\star = \max_{i \in [K]} \mu_i$$

Exploration-Exploitation Trade-off in MAB

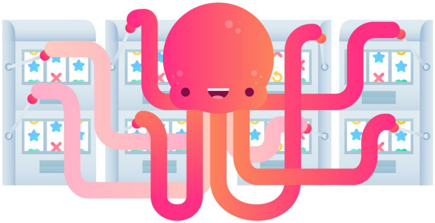
Should we pull arms that are less frequently tried in the past (i.e., explore), or should we commit to the current best arm (i.e., exploit)?

Exploration-Exploitation Trade-off in MAB

Should we pull arms that are less frequently tried in the past (i.e., explore), or should we commit to the current best arm (i.e., exploit)?

Let's try to only exploit and see what happens. We call this the **greedy algorithm.**

Greedy Algorithm



Algorithm:

- try each arm once
- commit to the one that has the highest observed reward

Greedy Algorithm

Algorithm:

- try each arm once
- commit to the one that has the highest observed reward

Problem: a (bad) arm with low μ_i may generate a high reward by chance, as we sample $r_i \sim \nu_i$ and it's i.i.d.

Consider two arms a_1, a_2: Reward dist for a_1: prob 60%: 1, else 0; for a_2: prob 40% 1, else 0. Now: a_1 is clearly better but with prob 16% we can observe (0, 1)

Greedy Algorithm: Lessons Learned

Trying the arm only once is not enough, since our sampled reward might be far from the mean

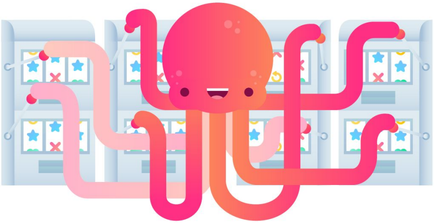
We can, however:

1. Try each arm multiple times
2. Compute the empirical mean of each arm
3. Commit to the arm with the highest empirical mean

Explore & Commit Algorithm

1. Set N to a fixed value, T >> K and K being the number of arms
2. For k = 1, ..., K: **(explore)**
   - pull arm *k* for N times
   - observe the set $\{r_i\}_{i=1}^{N} \sim \nu_i$
   - compute the empirical mean $\hat{\mu}_k = \sum_{i=1}^{N} r_i / N$
3. For t = NK, ..., T: **(commit)**
   - pull the best empirical arm

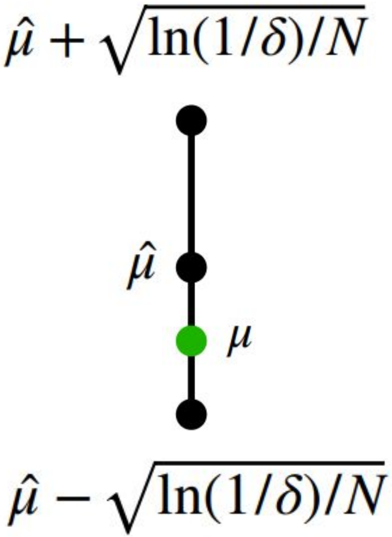
$$I_t = \arg\max_{i \in [K]} \hat{\mu}_i$$



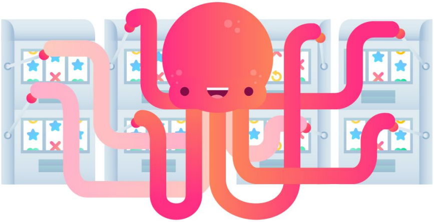
Hoeffding Inequality

Do we have a confidence interval on our empirical mean? <u>During exploration, for each arm</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

$$\left| \sum_{i=1}^{N} r_i/N - \mu_i \right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$



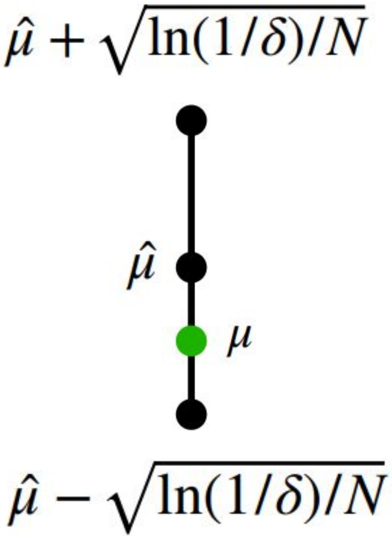
Hoeffding Inequality



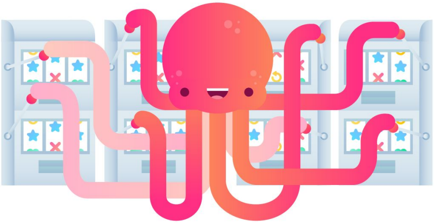
Do we have a confidence interval on our empirical mean? <u>During exploration, for each arm</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

$$\left|\boxed{\sum_{i=1}^{N} r_i/N} - \mu_i\right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$

our estimate

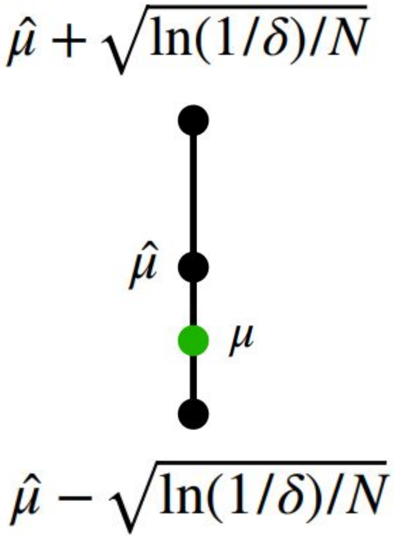


Hoeffding Inequality



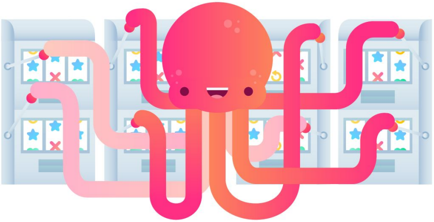
Do we have a confidence interval on our empirical mean? <u>During exploration, for each arm</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

$$\left| \sum_{i=1}^{N} r_i/N - \mu_i \right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$



e.g., δ = 0.01, confidence bound holds with probability 99%

Hoeffding Inequality



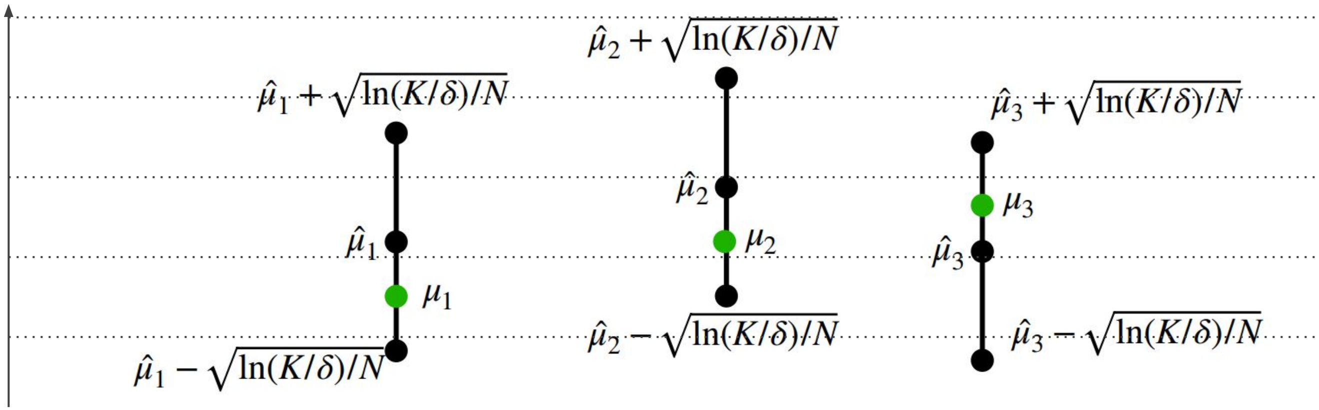
Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

$$\left| \sum_{i=1}^{N} r_i / N - \mu_i \right| \leq O\left(\sqrt{\frac{\ln(K/\delta)}{N}} \right)$$

K is the union of the k-th, different from the 1 used before

Hoeffding Inequality

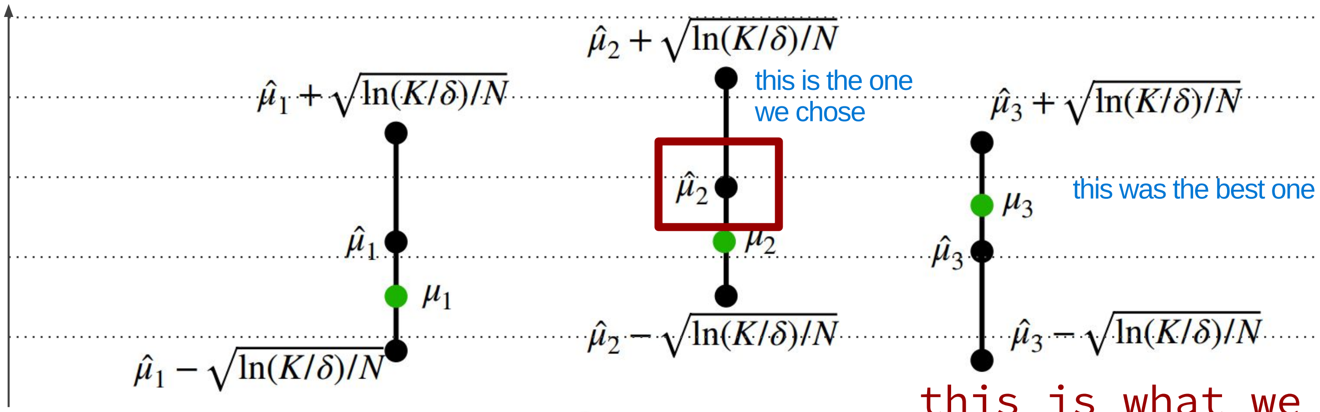
Do we have a confidence interval on our empirical mean? During exploitation, for all arms, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:



SAPIENZA
UNIVERSITÀ DI ROMA

Hoeffding Inequality

Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

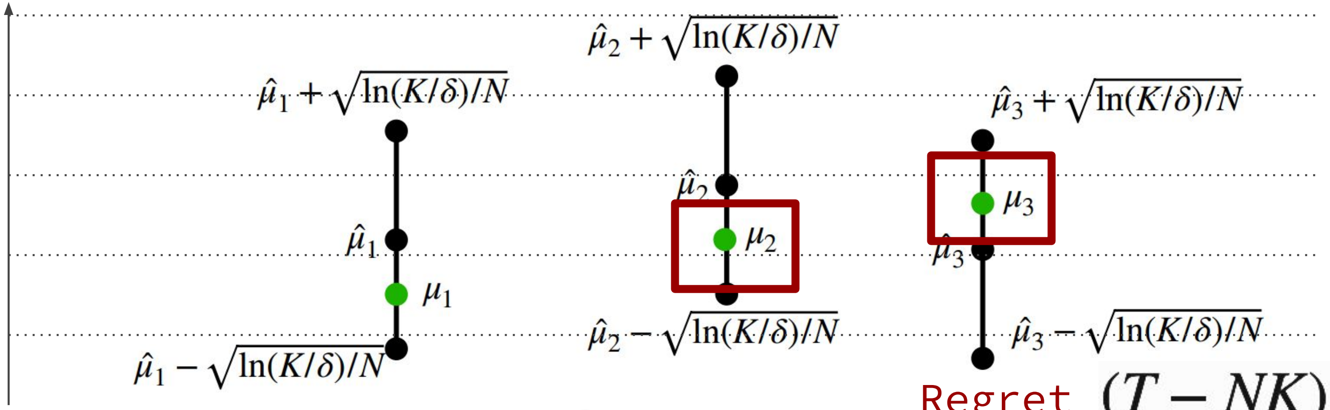


this is what we would pick

SAPIENZA UNIVERSITÀ DI ROMA

Hoeffding Inequality - Exploitation Regret

Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

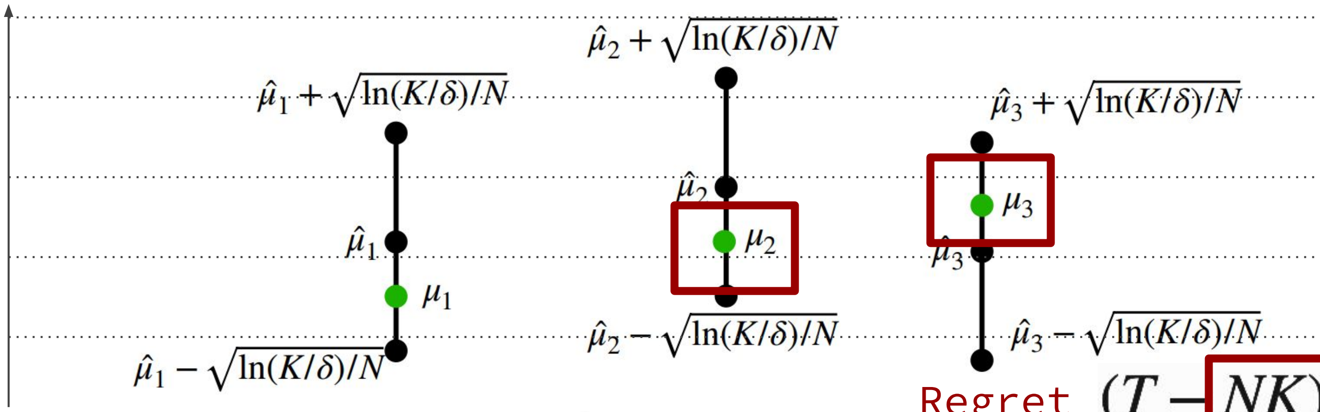


Regret $(T - NK)(\mu_3 - \mu_2)$

$$\text{Regret}_T = T\mu^\star - \sum_{t=0}^{T-1} \mu_{I_t}$$

Hoeffding Inequality - Exploitation Regret

Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

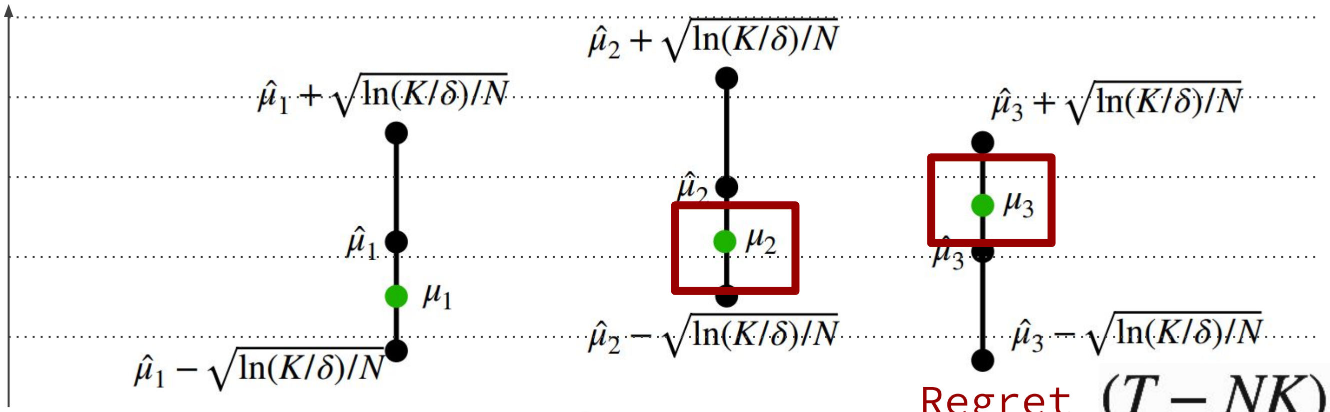


Exploration steps

Regret $(T - \boxed{NK})(\mu_3 - \mu_2)$

Hoeffding Inequality - Exploitation Regret

Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

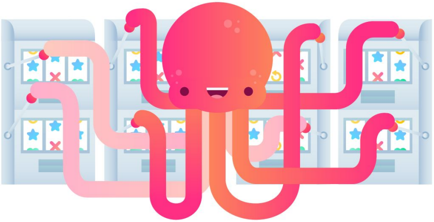


This difference is small, less of the confidence ranges

Regret $(T - NK)\boxed{(\mu_3 - \mu_2)}$

SAPIENZA
UNIVERSITÀ DI ROMA

Exploration Regret Calculation



- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

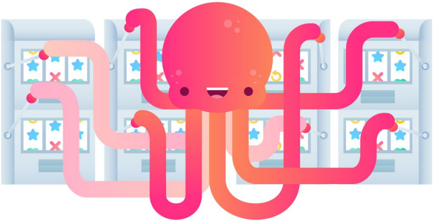
$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Trying all of the k-th arm N times
The best arm will have reward 1
Assuming all the other arms have 0.
This is the worst case

Worst possible regret in exploration: $\text{Regret}_{explore} \leq \boxed{N(K-1)} \leq NK$

We are trying all arms, including bad ones: maximum per-round regret is 1, as reward is in [0, 1]

Exploration Regret Calculation



- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Worst possible regret in exploration: $\text{Regret}_{explore} \leq N\boxed{(K-1)} \leq NK$

one arm is actually optimal

Exploitation Regret Calculation

- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Worst possible regret in exploitation: $\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right)$

$$\mu_{I^\star} - \mu_{\hat{I}} \leq \left[\hat{\mu}_{I^\star} + \sqrt{\ln(K/\delta)/N}\right] - \left[\hat{\mu}_{\hat{I}} - \sqrt{\ln(K/\delta)/N}\right] = \boxed{\hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}}} + 2\sqrt{\ln(K/\delta)/N} \leq 2\sqrt{\ln(K/\delta)/N}$$

rephrasing everything with the empirical mean, that is known

0 or smaller than 0, as I^ is the highest value according to our estimate

Regret Calculation

- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

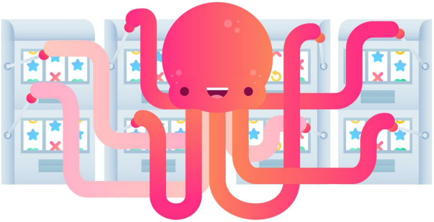
Worst possible regret in exploitation: $\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right)$

$$\mu_{I^\star} - \mu_{\hat{I}} \leq \left[\hat{\mu}_{I^\star} + \sqrt{\ln(K/\delta)/N}\right] - \left[\hat{\mu}_{\hat{I}} - \sqrt{\ln(K/\delta)/N}\right] = \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + 2\sqrt{\ln(K/\delta)/N} \leq \boxed{2\sqrt{\ln(K/\delta)/N}}$$

This regret is at worst twice our confidence interval

SAPIENZA
UNIVERSITÀ DI ROMA

Regret Calculation



- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Worst possible regret in exploitation: $\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right)$

$$\leq 2T\sqrt{\frac{\ln(K/\delta)}{N}}$$

SAPIENZA
UNIVERSITÀ DI ROMA

Regret Calculation

- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Total regret: $\text{Regret}_T = \text{Regret}_{explore} + \text{Regret}_{exploit} \leq NK + 2T\sqrt{\frac{\ln(K/\delta)}{N}}$

This regret is at worst twice our confidence interval

Regret Calculation

- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

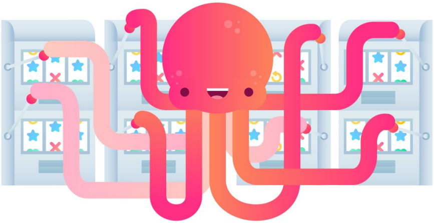
- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Total regret: $\text{Regret}_T = \text{Regret}_{explore} + \text{Regret}_{exploit} \leq NK + 2T\sqrt{\frac{\ln(K/\delta)}{N}}$

To minimize our regret, we want to optimize N: take the gradient of the regret, set it to 0, solve for N

Regret Calculation



- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^\star = \arg\max_{i \in [K]} \mu_i$$

Total regret: $\mathrm{Regret}_T = \mathrm{Regret}_{explore} + \mathrm{Regret}_{exploit} \leq NK + 2T\sqrt{\frac{\ln(K/\delta)}{N}}$

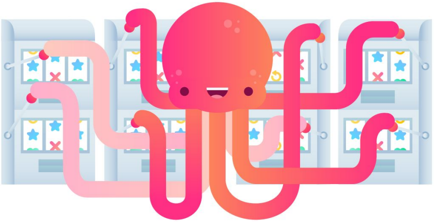
To minimize our regret, we want to optimize N: take the gradient of the regret, set it to 0, solve for N

speed 2/3 at which the regret is decaying

$$N = \left(\frac{T\sqrt{\ln(K/\delta)}}{2K}\right)^{2/3}$$

SAPIENZA
UNIVERSITÀ DI ROMA

Regret Calculation



- Empirical best arm:

$$\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$$

- Best arm:

$$I^{\star} = \arg\max_{i \in [K]} \mu_i$$

Total regret: $\text{Regret}_T = \text{Regret}_{explore} + \text{Regret}_{exploit} \leq NK + 2T\sqrt{\frac{\ln(K/\delta)}{N}}$

$$N = \left(\frac{T\sqrt{\ln(K/\delta)}}{2K}\right)^{2/3} \qquad \text{Regret}_T \leq O\left(T^{2/3}K^{1/3} \cdot \ln^{1/3}(K/\delta)\right)$$

Approaches 0 as T goes to infinite

SAPIENZA
UNIVERSITÀ DI ROMA

Regret Decaying

The decaying rate of the regret using the explore & commit algorithm is kind of slow ($T^{2/3}$). Can we get something faster, like $O(\sqrt{T})$?

Regret Decaying

The decaying rate of the regret using the explore & commit algorithm is kind of slow ($T^{2/3}$). Can we get something faster, like $O(\sqrt{T})$?

$O(\sqrt{T})$ is actually the minimum we can get as it is a lower bound (no algorithm ever will be faster than this)

Regret Decaying

The decaying rate of the regret using the explore & commit algorithm is kind of slow ($T^{2/3}$). Can we get something faster, like $O(\sqrt{T})$?

$O(\sqrt{T})$ is actually the minimum we can get as it is a lower bound (no algorithm ever will be faster than this)

Let's try to design a new algorithm

Statistics to Maintain & Confidence

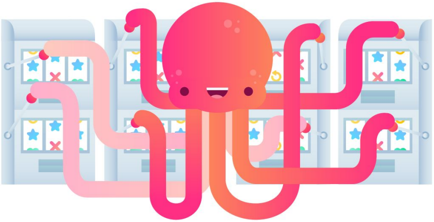
Let’s write a list of generic statistics that we need to maintain in order to compute our confidence bounds and the regret

- # of times we have tried arm i $$N_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\}$$ 1 is indicator func

- empirical mean so far $$\hat{\mu}_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\} r_\tau / N_t(i)$$

Confidence with probability 1-δ: $$|\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$$

Statistics to Maintain & Confidence



Let's write a list of generic statistics that we need to maintain in order to compute our confidence bounds and the regret

- # of times we have tried arm i $N_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\}$
- empirical mean so far $\hat{\mu}_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\} r_\tau / N_t(i)$

Confidence with probability 1-δ: $|\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\frac{\ln(K\boxed{T}/\delta)}{N_t(i)}}$

union bound over all T

Statistics to Maintain & Confidence

Let's write a list of generic statistics that we need to maintain in order to compute our confidence bounds and the regret

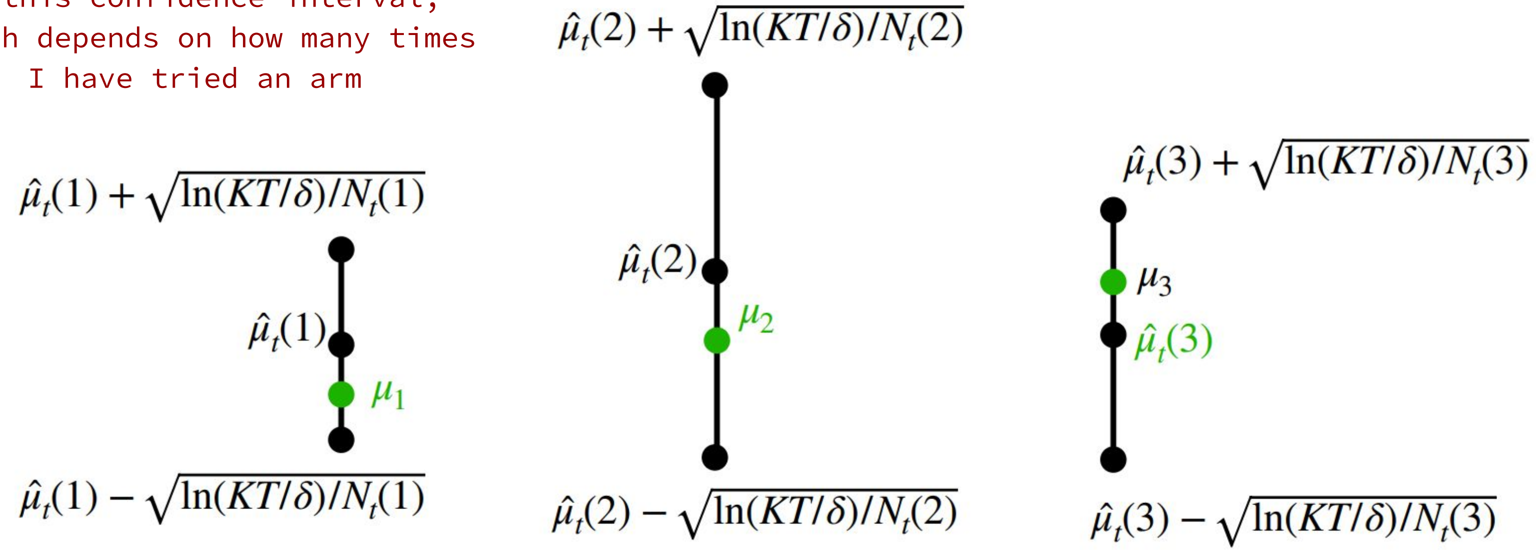
- # of times we have tried arm i $N_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\}$
- empirical mean so far $\hat{\mu}_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\} r_\tau / N_t(i)$

this is a confidence interval for all iterations and all arms!

Confidence with probability 1-δ: $|\hat{\mu}_t(i) - \mu_i| \leq \boxed{\sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}}$

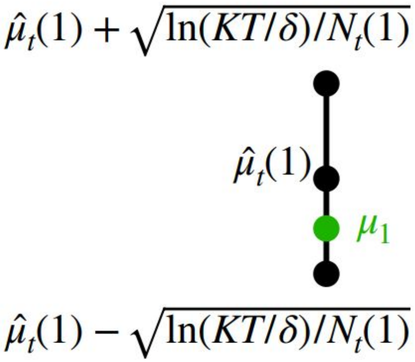
Optimism in the Face of Uncertainty

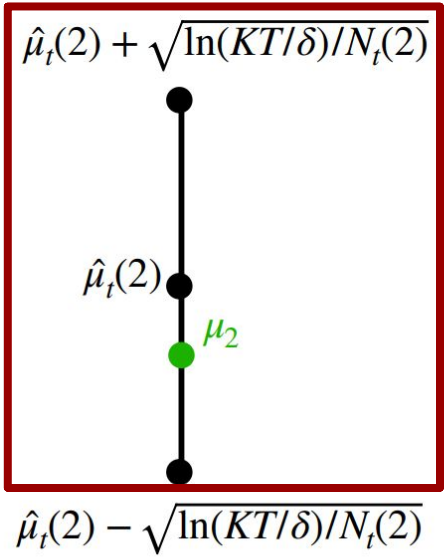
In this confidence interval, length depends on how many times I have tried an arm

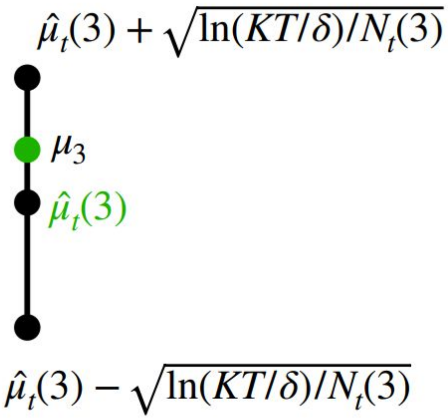


Optimism in the Face of Uncertainty

The length of the confidence of this arm is higher because I did not try arm 2 as many times as arm 1 and 3



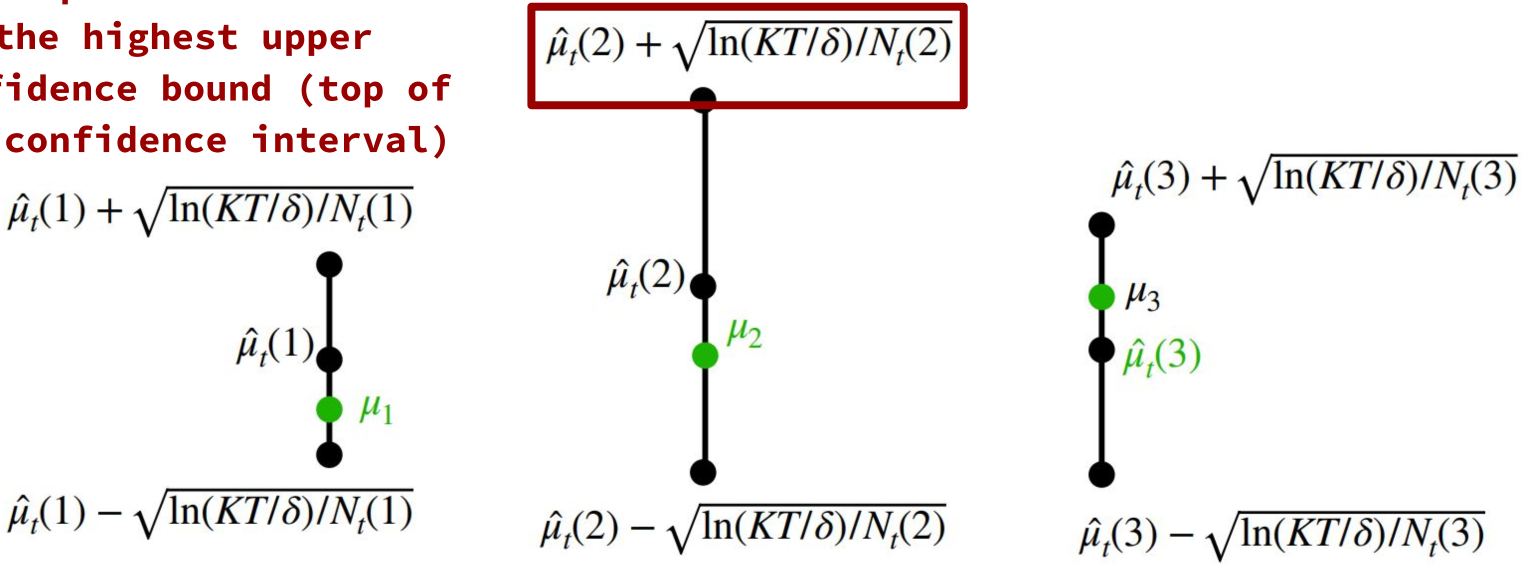




Optimism in the Face of Uncertainty

For exploration: Either we have the highest because we didn't explore enough so the bounds high, so we explore and understand better if it's good -> we are optimistic
For exploitation: The bounds are already shrinked enough (second added is small) but the mean is very high (first addend is high)

Let's pick the arm with the highest upper confidence bound (top of the confidence interval)



Optimism in the Face of Uncertainty

example

	t = 0	t = 1	t = 2
I_1	N_0(1)	N_1(1)	N_2(1)
I_2	N_0(2)	N_1(2)	N_2(2)
I_3	N_0(3)	N_1(3)	N_2(3)

Ad mu_0(1), mu_0(2) ... (empirical)

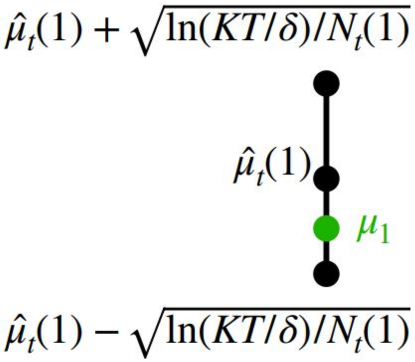
Let's say we ger r = 1 for I_2 and 0 for all the others at time t

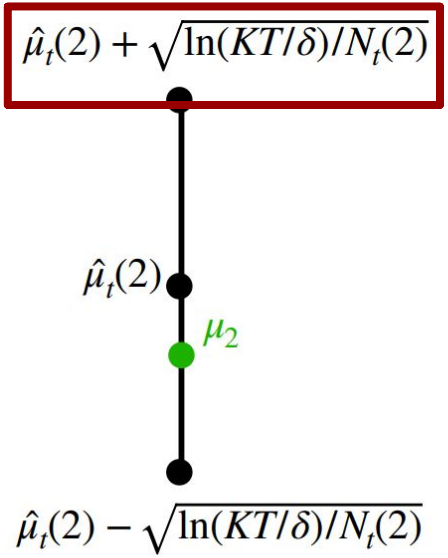
N_i(j) will be 1 for each j and mu_0(2) = 1 while the other mu = 0

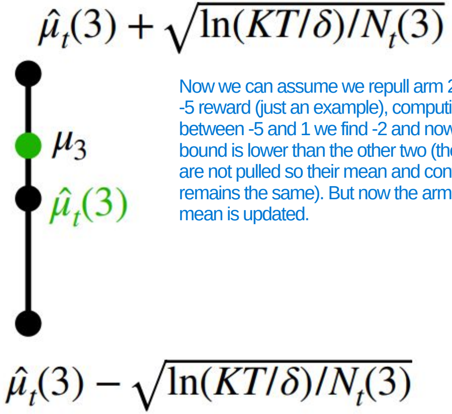
Now let's compute the bounds

k = 3, T = number of total time you want to interact (let's say 100), delta whatever

We are optimistic about the fact that the true mean actually corresponds to the upper confidence bound







Now we can assume we repull arm 2 and it has -5 reward (just an example), computing the mean between -5 and 1 we find -2 and now the confidence bound is lower than the other two (the other two arm are not pulled so their mean and confidence bound remains the same). But now the arm 2 confidence and mean is updated.

UCB Algorithm

- For the first K iterations, pull each arm once
- For t = K, ..., T:
  - pick the action with the highest upper confidence bound

$$I_t = \arg\max_{i \in [K]} \left(\hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}} \right)$$

  - update statistics

UCB Algorithm

- For the first K iterations, pull each arm once
- For t = K, ..., T:
  - pick the action with the highest upper confidence bound

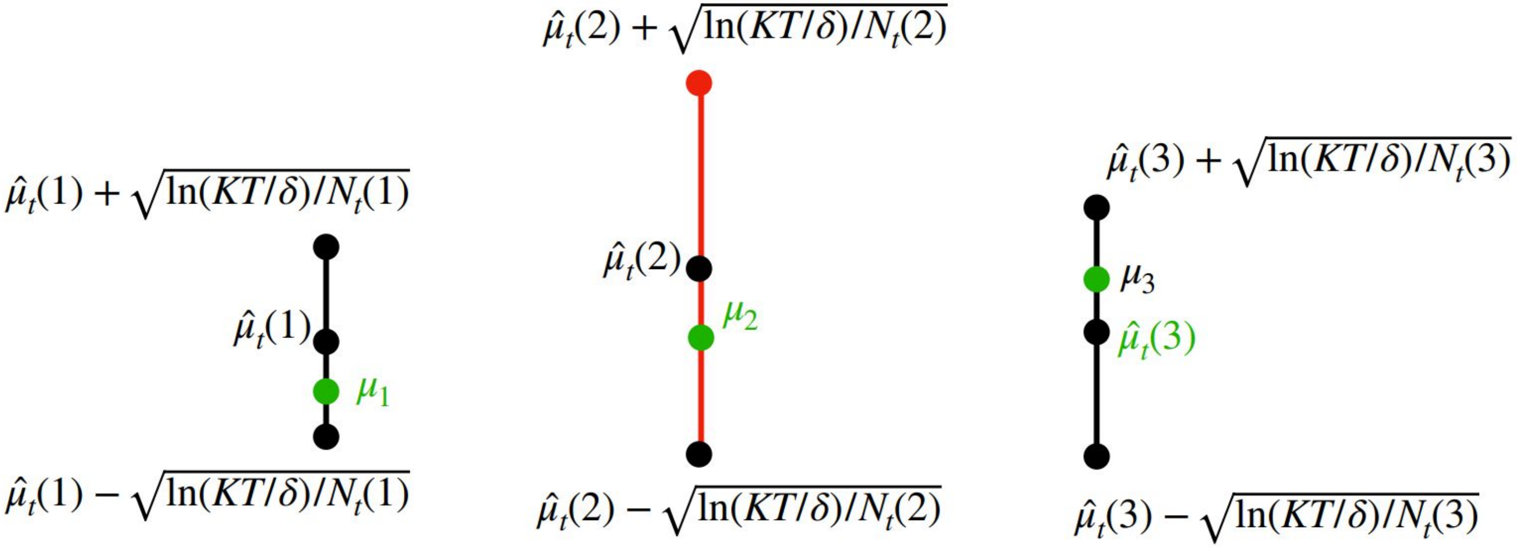
$$I_t = \arg\max_{i \in [K]} \left(\hat{\mu}_t(i) + \boxed{\sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}} \right)$$

  - update statistics

Reward bonus is high if we did not try action many times: exploration

UCB Algorithm: Intuition

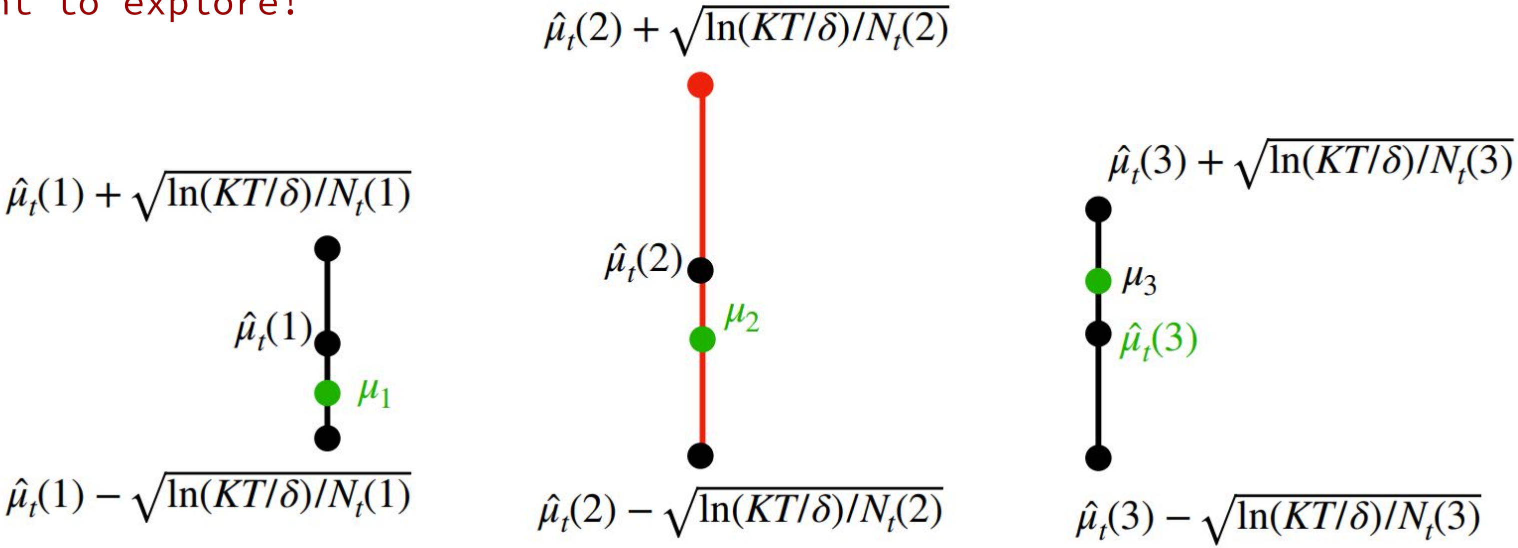
Case 1: large confidence interval, not tried many times (high uncertainty)



UCB Algorithm: Intuition

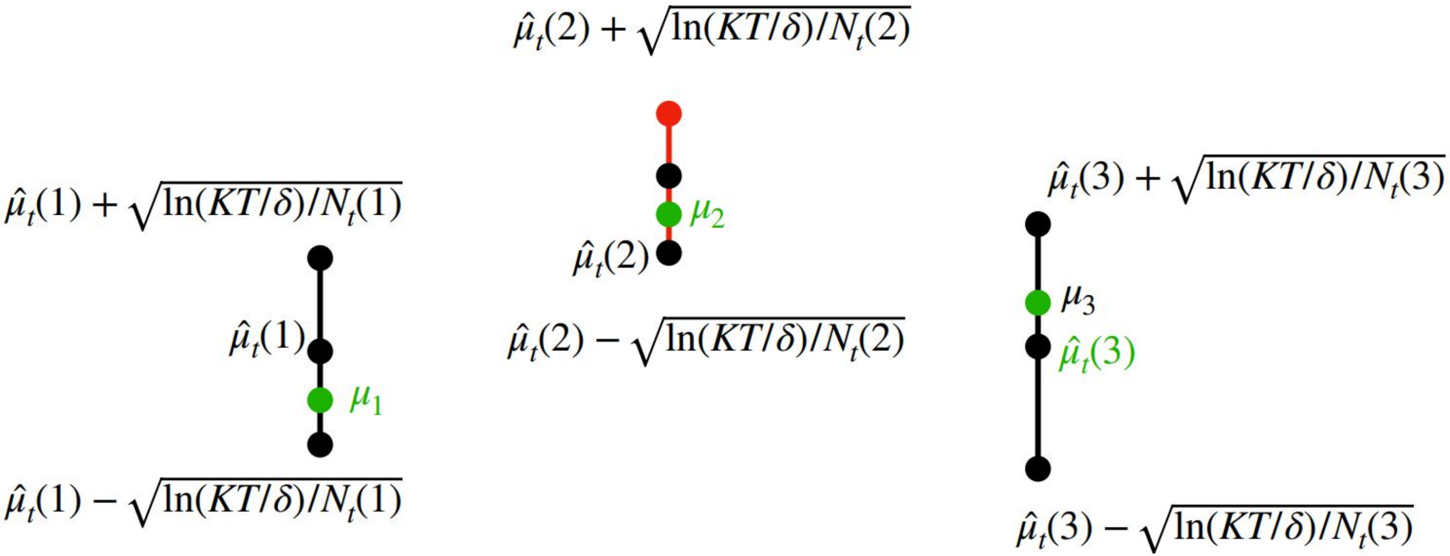
Case 1: large confidence interval, not tried many times (high uncertainty)

We want to explore!



UCB Algorithm: Intuition

Case 2: small confidence interval, good arm: true mean is high

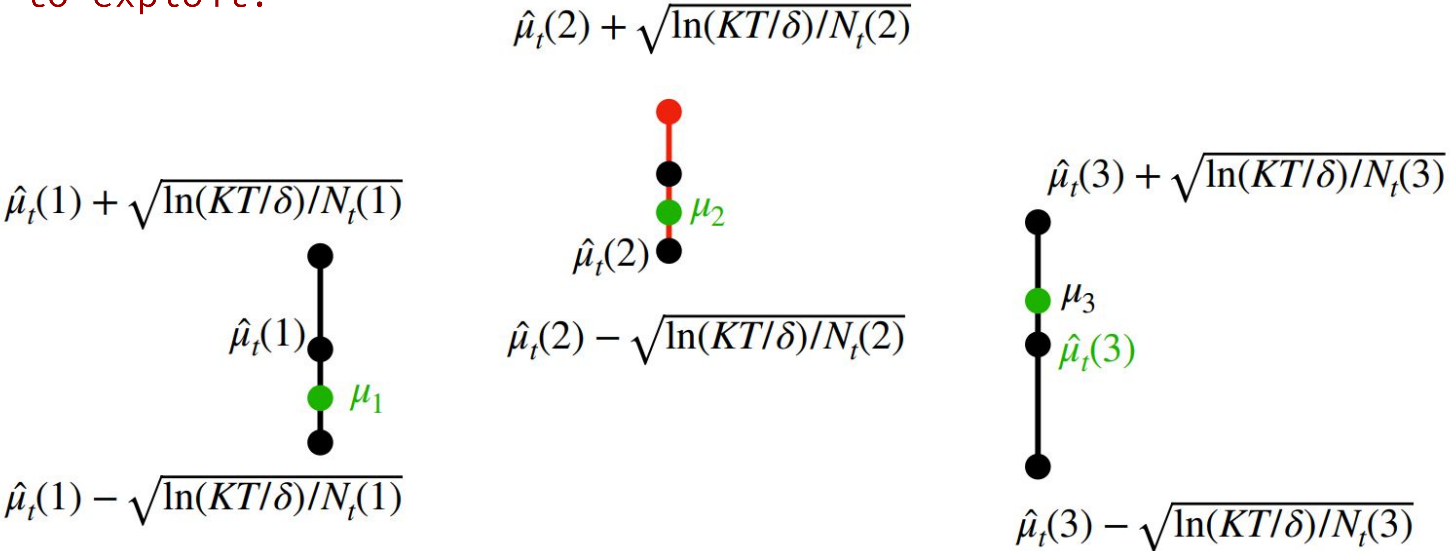


SAPIENZA UNIVERSITÀ DI ROMA

UCB Algorithm: Intuition

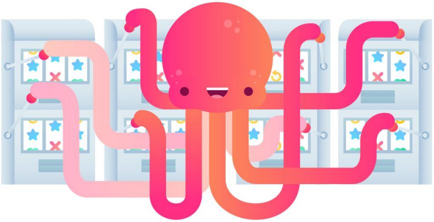
Case 2: small confidence interval, good arm: true mean is high

We want to exploit!



SAPIENZA
UNIVERSITÀ DI ROMA

UCB Algorithm: Regret-at-t

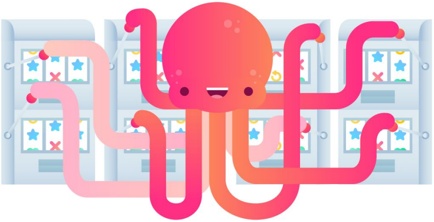


$$I^\star = \arg\max_{i \in [K]} \mu_i$$

$$I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t} \le \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \le 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

UCB Algorithm: Regret-at-t

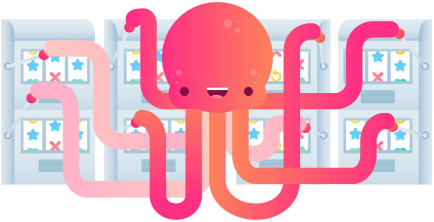


$$\boxed{I^\star = \arg\max_{i \in [K]} \mu_i} \qquad \boxed{I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}}$$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t} \leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

Case 1: N_t is small. We have regret but we explore (select I_t at iteration t)

UCB Algorithm: Regret-at-t



$$I^\star = \arg\max_{i \in [K]} \mu_i$$

$$I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t} \leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

Case 2: N_t is large. We exploit (select I_t at iteration t) and regret is small

UCB Algorithm: Regret

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t} \le \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \le 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\text{Regret}_T = \sum_{t=0}^{T-1}\left(\mu^\star - \mu_{I_t}\right) \le \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \le 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1}\sqrt{\frac{1}{N_t(I_t)}}$$

UCB Algorithm: Regret

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t} \le \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \le 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\text{Regret}_T = \sum_{t=0}^{T-1}\left(\mu^\star - \mu_{I_t}\right) \le \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \le 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1}\sqrt{\frac{1}{N_t(I_t)}}$$

dragging outside since it's constant w.r.t t

$$\sum_{t=0}^{T-1}\sqrt{\frac{1}{N_t(I_t)}} \le O\left(\sqrt{KT}\right) \longrightarrow$$ **With high probability** $\text{Regret}_T = \widetilde{O}\left(\sqrt{KT}\right)$

Proof of this inequality in the next slide

SAPIENZA UNIVERSITÀ DI ROMA

See also https://wensun.github.io/CS4789_data/UCB_note_new.pdf

UCB Algorithm: Regret

$$\sum_{t=0}^{T-1} \sqrt{1/n_t(I_t)} = \sum_{i=1}^{K} \sum_{t=0}^{T-1} \mathbf{1}[I_t = i]\sqrt{1/n_t(i)}$$

cluster pulled arms into K groups where the i-th group contains all steps where arm i is pulled

$$= \sum_{i=1}^{K} \sum_{t=1}^{n_T(i)} \sqrt{1/t}$$

$n_T(i)$ is the size of the specific group

$$\leq \sum_{i=1}^{K} \sqrt{n_T(i)}$$

uses the trick that $\sum_{i=1}^{N} 1/\sqrt{i} \leq \sqrt{N}$

$$K\frac{1}{K}\sum_{i=1}^{K}\sqrt{n_T(i)} \leq K\sqrt{\frac{1}{K}\sum_{i=1}^{K} n_T(i)} = K\sqrt{T/K} \longrightarrow \sqrt{KT}$$

All $n_T(i)$ sum to T + Jensen's inequality + K = sqrt(K^2)

See also
https://wensun.github.io/CS4789_data/UCB_note_new.pdf

SAPIENZA
UNIVERSITÀ DI ROMA