Exploration in RL: Regret and Multi-Armed Bandits

Roberto Capobianco

SAPIENZA
UNIVERSITÀ DI ROMA

Adapted from Wen Sun's slides

Recap

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what’s their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = ??$$

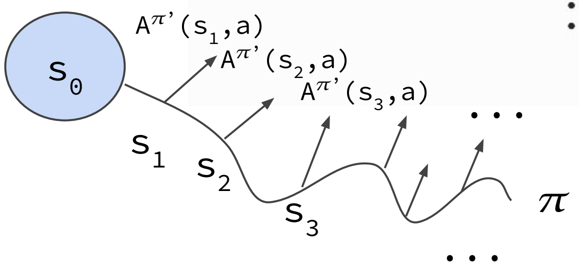
Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what's their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

$$:= \frac{1}{1-\gamma}\mathbb{E}_{\boxed{s\sim d^{\pi}_{s_0}}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\boxed{A^{\pi'}(s,a)}\right]$$

Average advantage value



Incremental Update of CPI

Oscillations are due to the distribution change induced by the policy

Can we design an update rule that does not change the distribution so much?

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

$$\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\|_1 \leq 2\alpha \longrightarrow \|d_{\mu}^{\pi^{t+1}}(\cdot) - d_{\mu}^{\pi^{t}}(\cdot)\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

Problem of CPI

I now need to retain all the old policies in memory: what if they are all large neural networks?

Let's use KL-Divergence

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s, a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

$$\boxed{\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots}$$

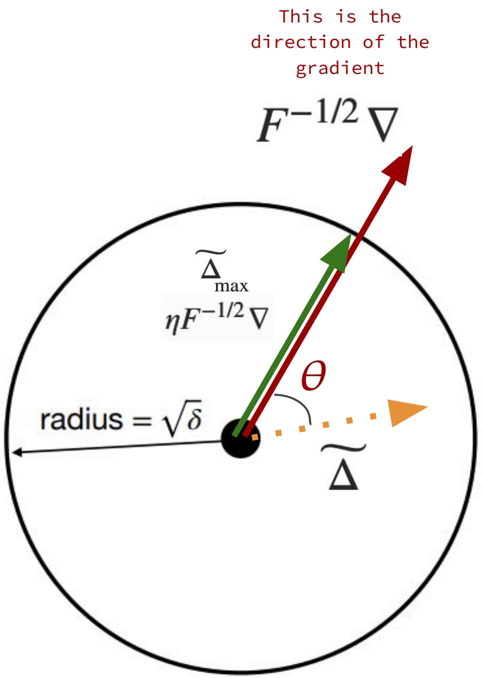
Simplified Trust-Region Formulation

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

$$\boxed{\begin{aligned} &\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) \\ &\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta \end{aligned}}$$

Simplified Trust-Region Solution



$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\top}\widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^{\top}\widetilde{\Delta} \leq \delta$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$$\|\eta F^{-1/2}\nabla\|_2 = \sqrt{\delta} \quad \Rightarrow \eta = \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}$$

$$\widetilde{\Delta}_{max} := \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}F^{-1/2}\nabla$$

$$\boxed{\Delta_{max} := \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}F^{-1}\nabla}$$

SAPIENZA UNIVERSITÀ DI ROMA

Credits: Wen Sun

TRPO: Line Search

Due to the quadratic approximation, the KL constraint might be violated: we solve this by doing a simple line search

for $j = 0, 1, 2, \ldots, L$ **do**
  Compute proposed update $\theta = \theta_k + \alpha^j \Delta_k$
  if $\mathcal{L}_{\theta_k}(\theta) \geq 0$ and $\bar{D}_{KL}(\theta || \theta_k) \leq \delta$ **then**
    accept the update and set $\theta_{k+1} = \theta_k + \alpha^j \Delta_k$
    break
  end if
end for

Natural Policy Gradient: Additional Comments

We want to keep two distributions close, but parameters can change a lot: learning rate (η) is very high if eigen-values of F are very small (as the matrix is inverted)

Generally, Natural PG moves faster than standard/plain PG

If we have many parameters, computing & inverting F is too heavy!

Extending TRPO: Proximal Policy Optimization

If we have many params, we can impose KL divergence as a regularization term and optimize (simply through SG Ascent)

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] \underbrace{-\lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a\,|\,s)\,|\,\pi_\theta(a\,|\,s)\right)\right]}_{\text{regularization}}$$

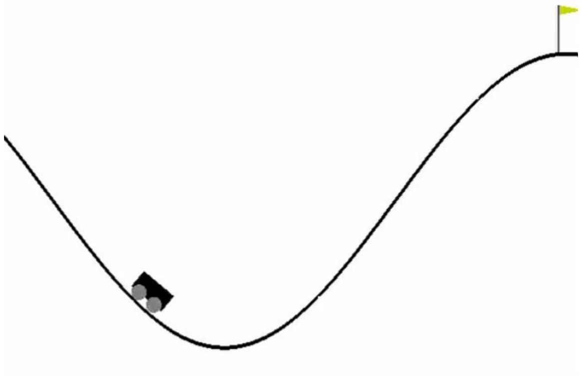
using importance weighting and expanding KL divergence through expectation

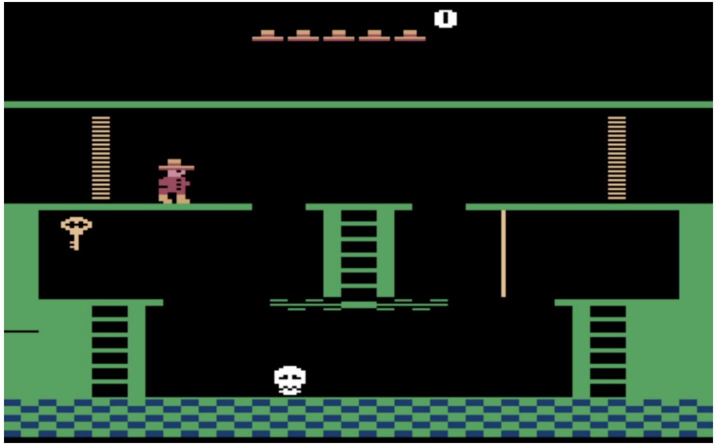
$$\ell(\theta) := \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_\theta(a\,|\,s)}{\pi_{\theta_t}(a\,|\,s)} A^{\pi_{\theta_t}}(s,a)\right] - \lambda \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\left[-\ln \pi_\theta(a\,|\,s)\right]$$

End Recap

Failure Mode of RL

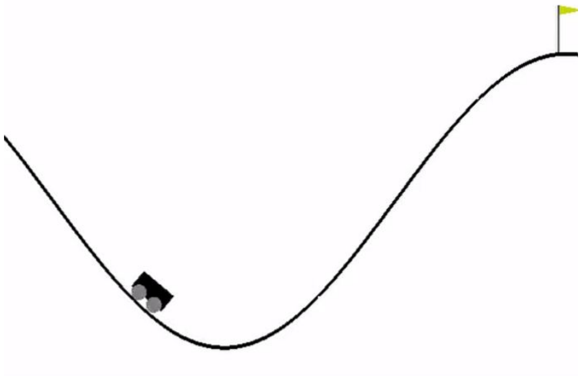
Sparse rewards (e.g., Mountain Car or Montezuma's Revenge) are a problem in RL: zero reward everywhere except in few states

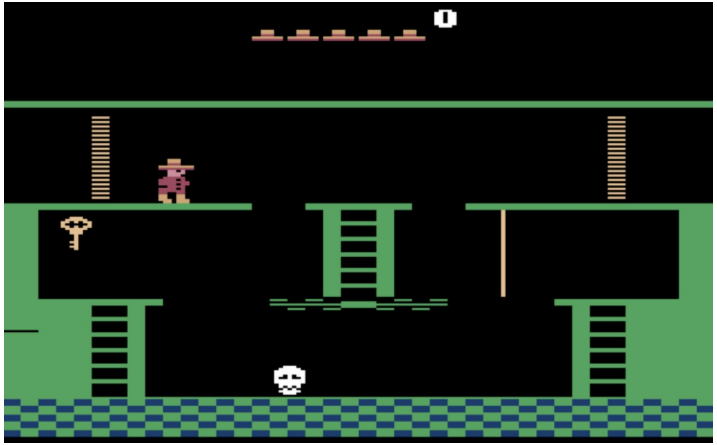




Failure Mode of RL

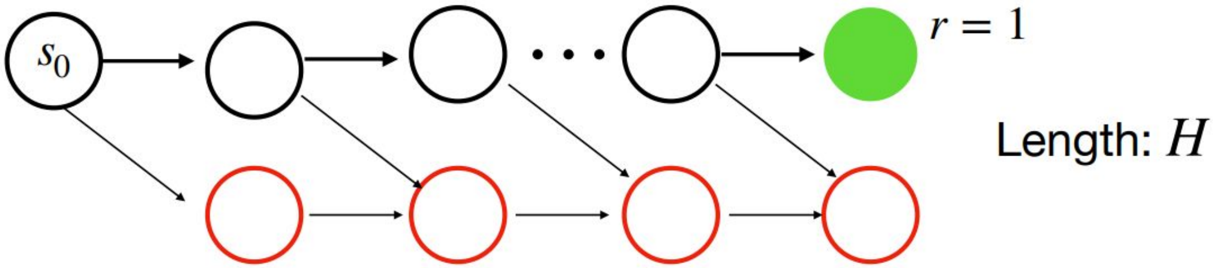
The probability of hitting those non-zero reward states is exponentially small!





Failure Mode of RL

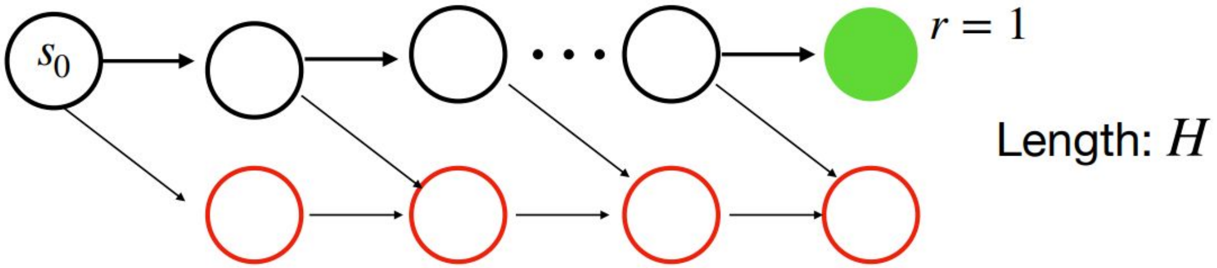
Consider the following MDP: two actions, zero reward everywhere, except in green state; one of the actions always leads to a dead state (can never get reward from there)



Failure Mode of RL

Consider the following MDP: two actions, zero reward everywhere, except in green state; one of the actions always leads to a dead state (can never get reward from there)

What is the probability of a random policy hitting the goal (i.e., getting to a reward)?



Exploration: the Big Pain of RL

We need to carefully and systematically explore (remember states we visited, and try to visit unexplored regions)

Exploration: the Big Pain of RL

We need to carefully and systematically explore (remember states we visited, and try to visit unexplored regions)

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)

Exploration: the Big Pain of RL

We need to carefully and systematically explore (remember states we visited, and try to visit unexplored regions)

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)

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

Multi-Armed Bandit

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

Multi-Armed Bandit

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

- One single state
- K different arms (think of them as actions): $a_1, \ldots, a_k$
- 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

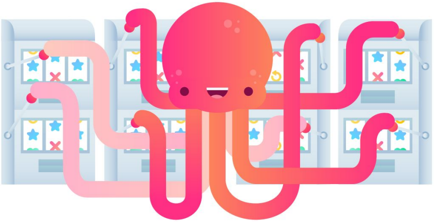
Multi-Armed Bandit: Example

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

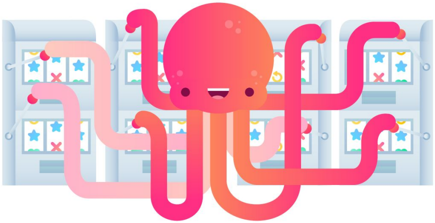


The interactive process that we deal with in MAB is the following:

For t = 0, ..., T-1:

1. Pull an arm I_t in {1, ..., 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

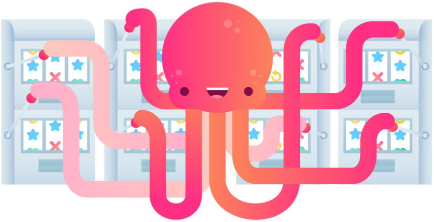
The interactive process that we deal with in MAB is the following:

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

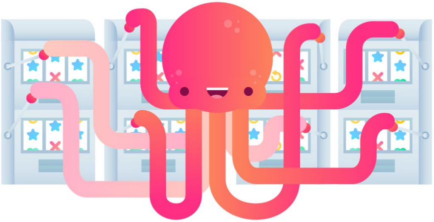
We want to minimize our **opportunity loss**, which is expressed in the form of the regret

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**

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

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

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

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

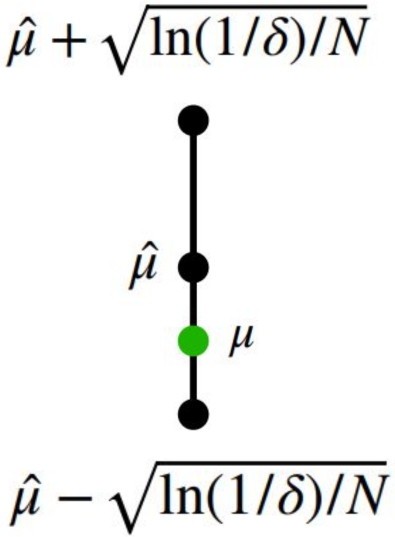
1. Set N = T/K, where T >> K and K is the number of arms
2. For k = 1, ..., K: **(explore)**
   - pull arm *k* for N times
   - observe the set $\{r_i\}_{i=1}^{N} \sim \nu_i$
   - compute the empirical mean $\hat{\mu}_k = \sum_{i=1}^{N} r_i / N$
3. For t = NK, ..., T: **(commit)**
   - pull the best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

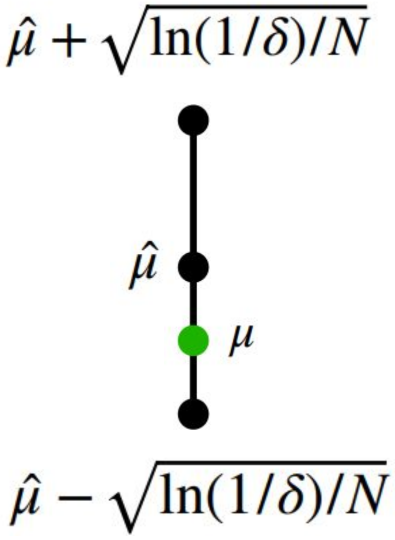
$$\left|\sum_{i=1}^{N} r_i/N - \mu_i\right| \le O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$



Hoeffding Inequality

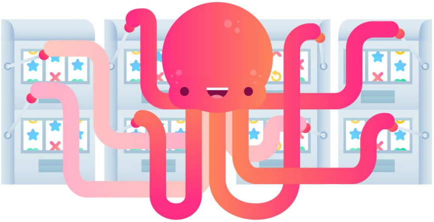
Do we have a confidence interval on our empirical mean? <u>During exploration, for each arm</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:

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



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

Hoeffding Inequality

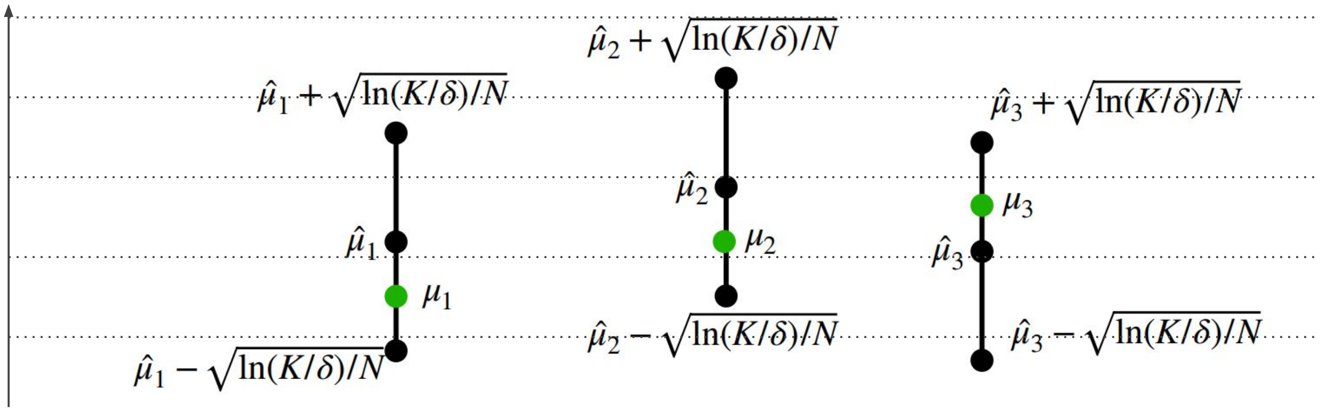


Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability $1-\delta$:

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

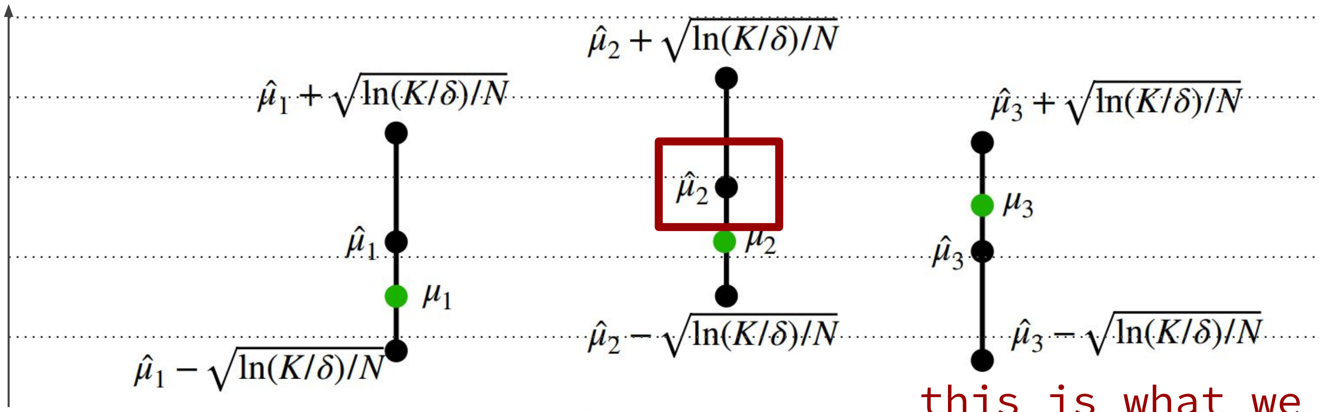
Hoeffding Inequality

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:



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:

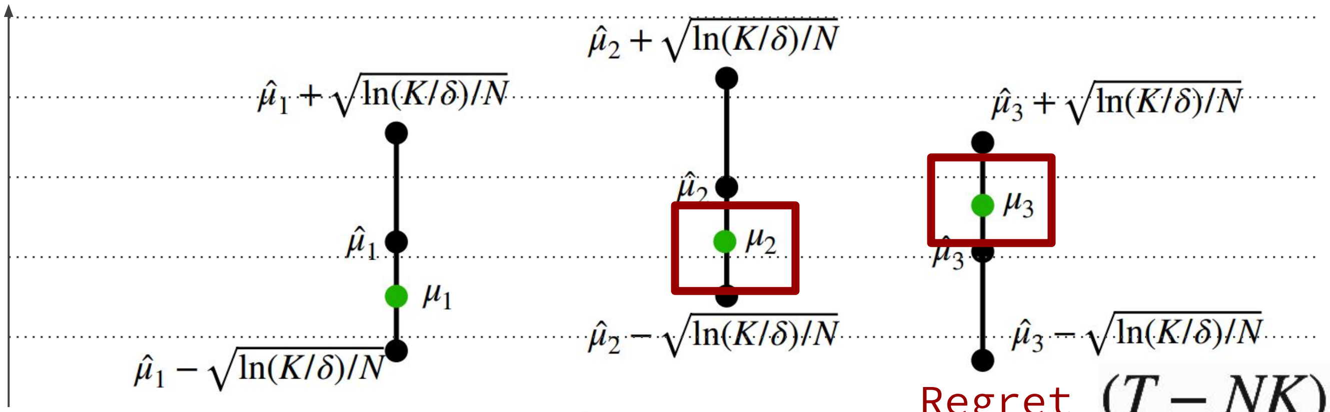


this is what we would pick

SAPIENZA
UNIVERSITÀ DI ROMA

Hoeffding Inequality

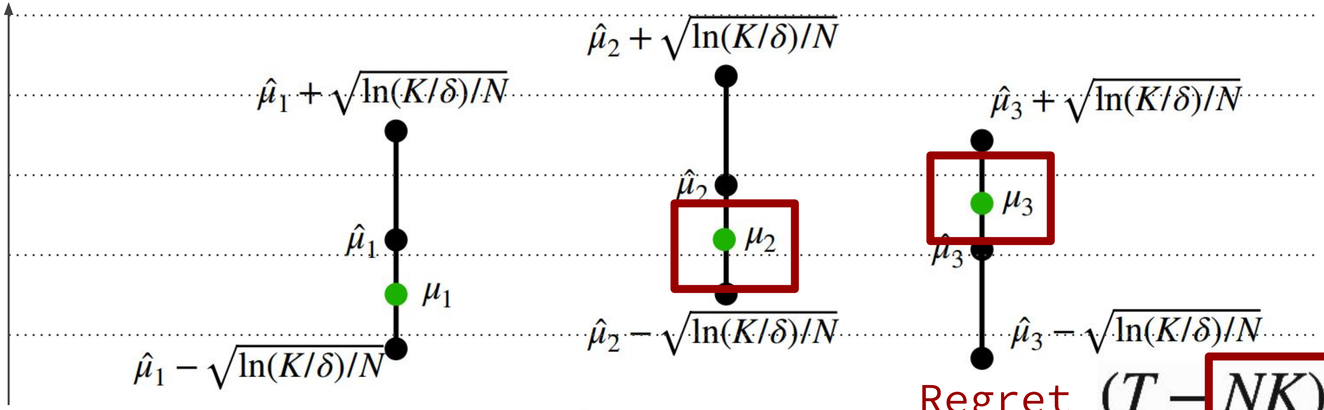
Do we have a confidence interval on our empirical mean? <u>During exploitation, for all arms</u>, given a distribution with mean μ and N i.i.d. samples, we have with probability 1-δ:



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

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:

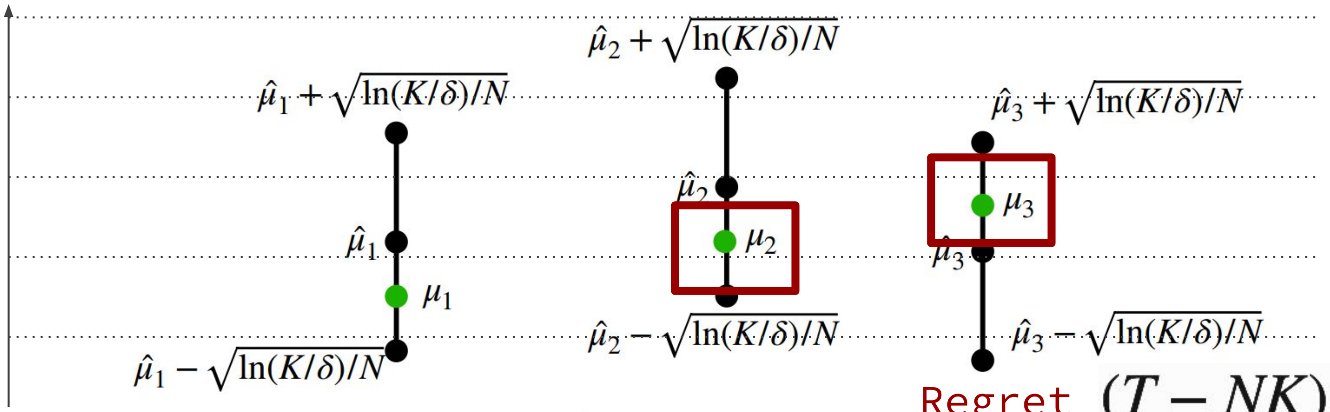


Exploration steps

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

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 difference is small, less of the confidence ranges

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

Regret Calculation

- Empirical best arm:

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

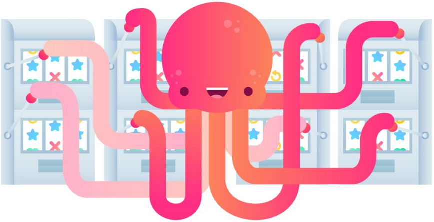
- Best arm:

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

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

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

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

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

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

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

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

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

This regret is at worst twice our confidence interval

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

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

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

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

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:

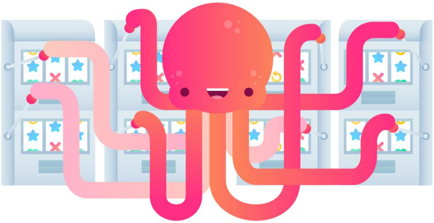
$$I^\star = \arg\max_{i \in [K]} \mu_i$$

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

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

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

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

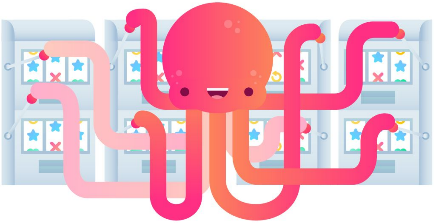
Approaches 0 as T goes to infinite

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 $O(\sqrt{T})$?

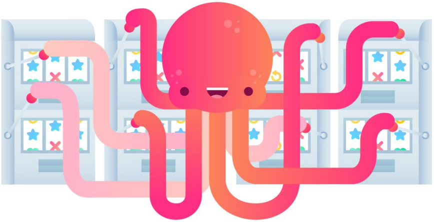
Regret Decaying



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

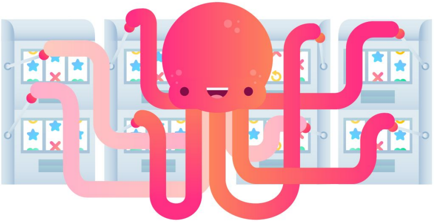
Statistics to Maintain & Confidence

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 $N_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\}$
- 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 $N_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\}$

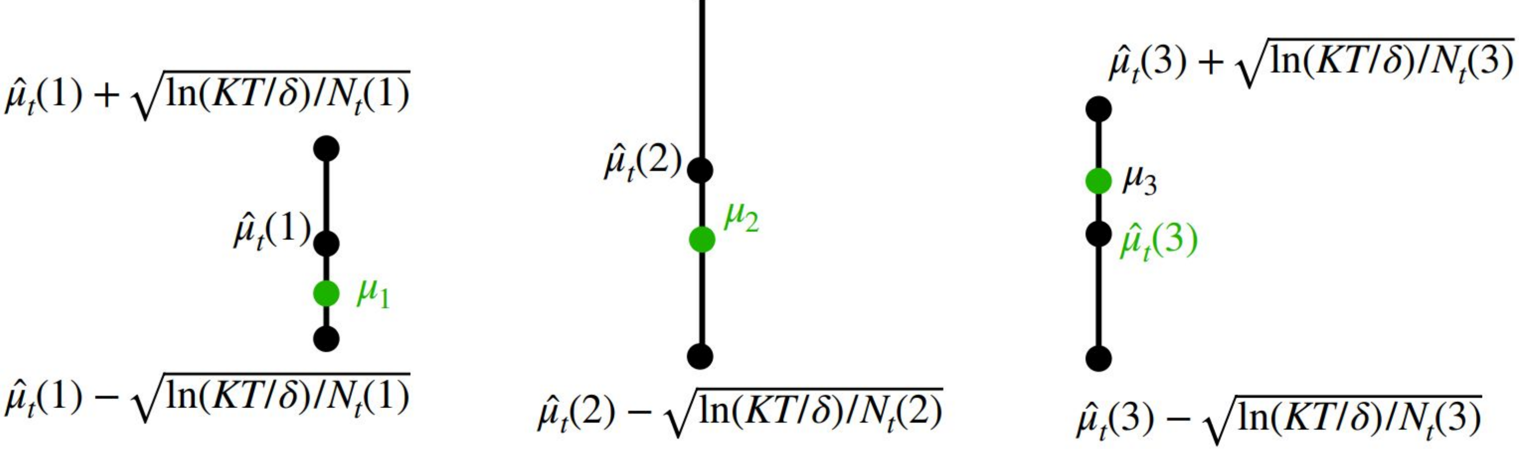
- empirical mean so far $\hat{\mu}_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\} r_\tau / N_t(i)$

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

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

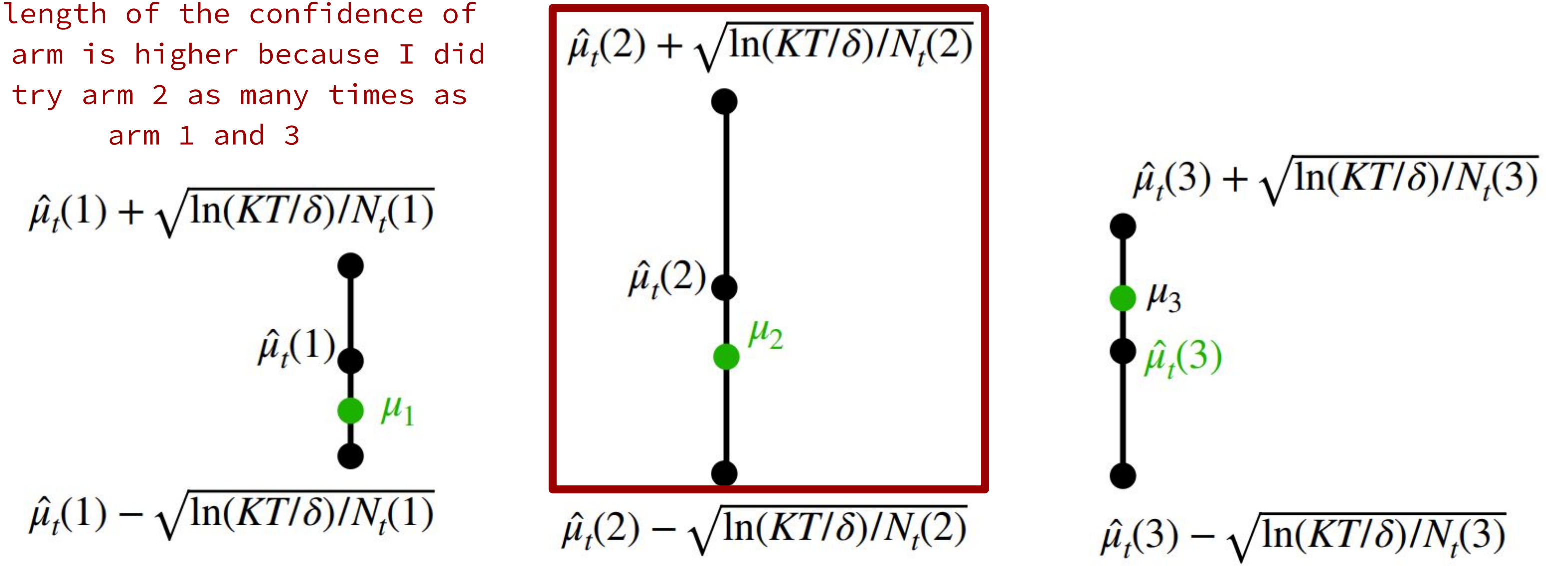
Optimism in the Face of Uncertainty

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



Optimism in the Face of Uncertainty

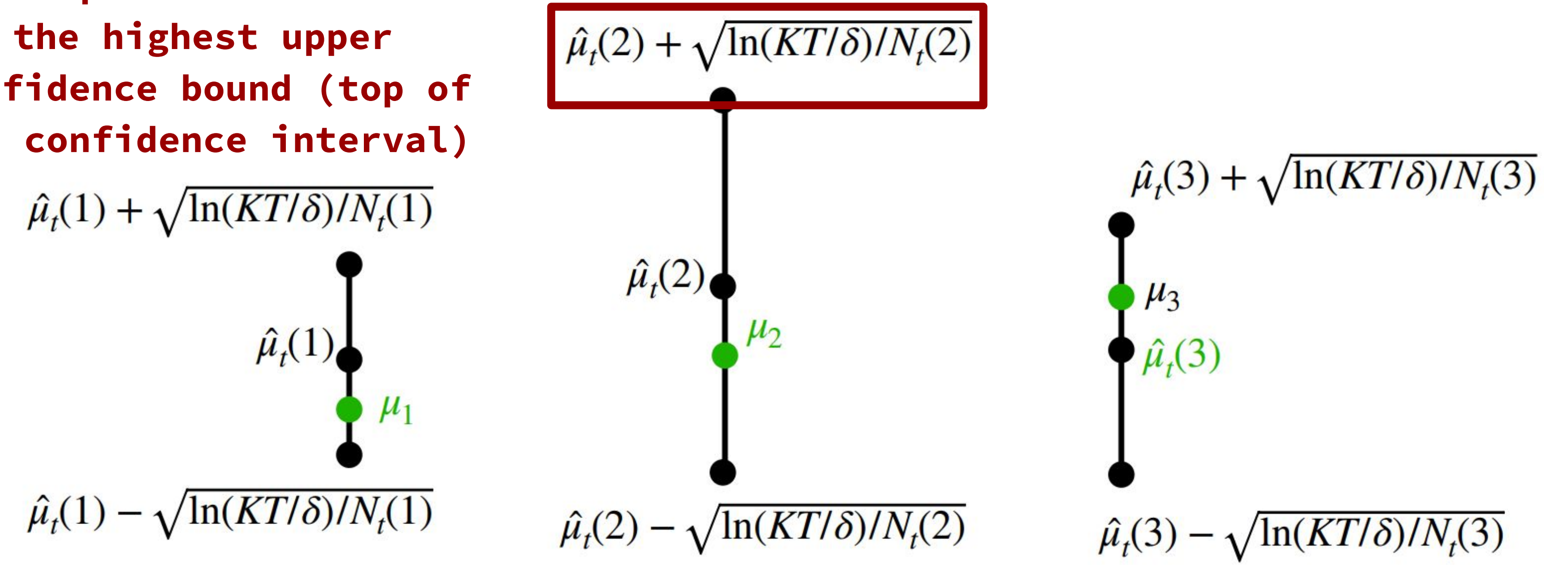
The length of the confidence of this arm is higher because I did not try arm 2 as many times as arm 1 and 3



SAPIENZA
UNIVERSITÀ DI ROMA

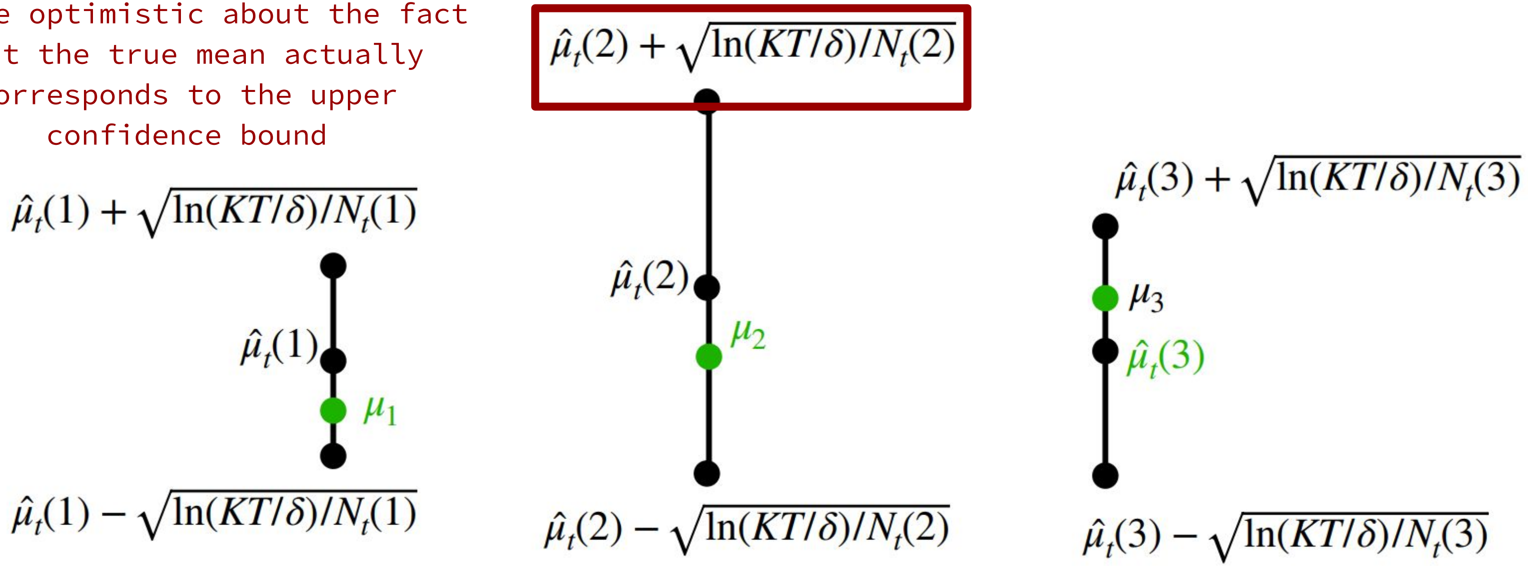
Optimism in the Face of Uncertainty

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

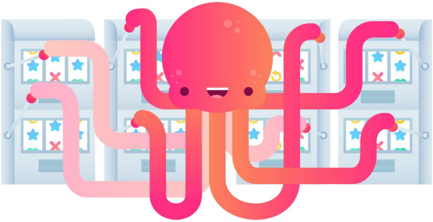


Optimism in the Face of Uncertainty

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



UCB Algorithm

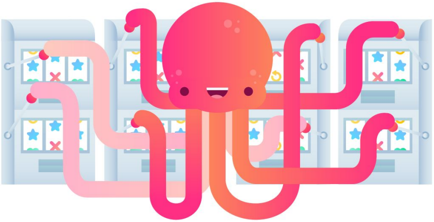


- For the first K iterations, pull each arm once
- For t = K, ..., T:
  - pick the action with the highest upper confidence bound

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

  - update statistics

UCB Algorithm



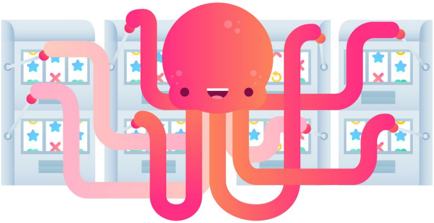
- For the first K iterations, pull each arm once
- For t = K, ..., T:
  - pick the action with the highest upper confidence bound

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

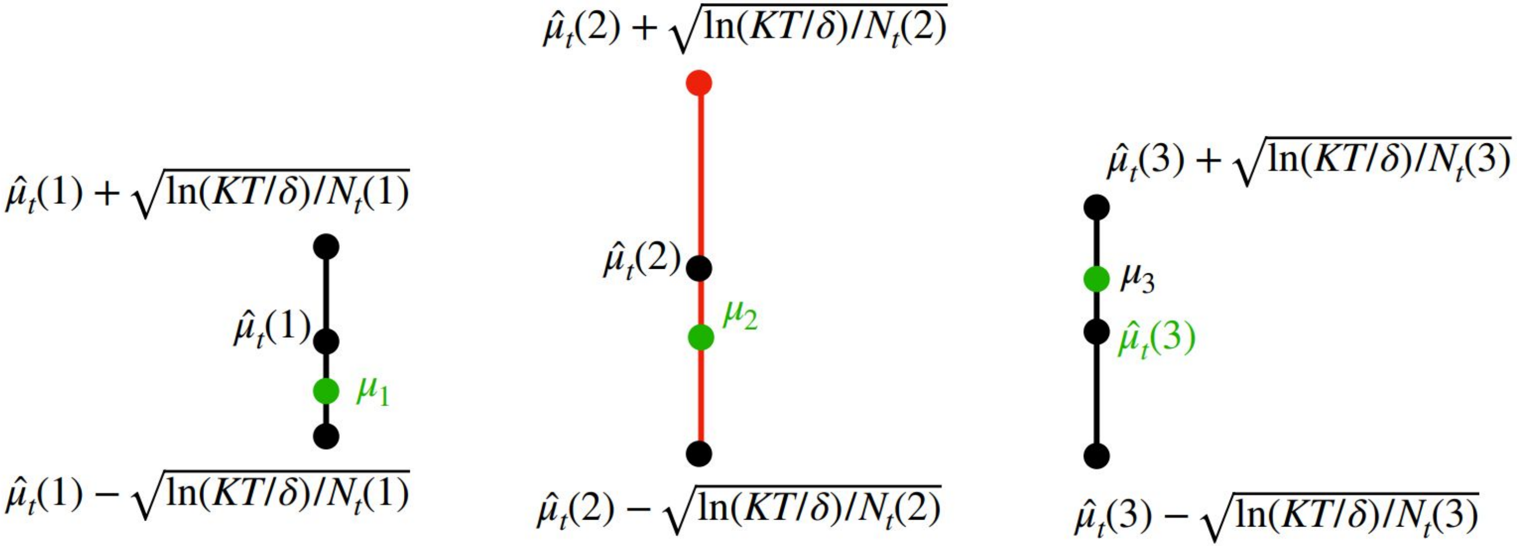
  - update statistics

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

UCB Algorithm: Intuition



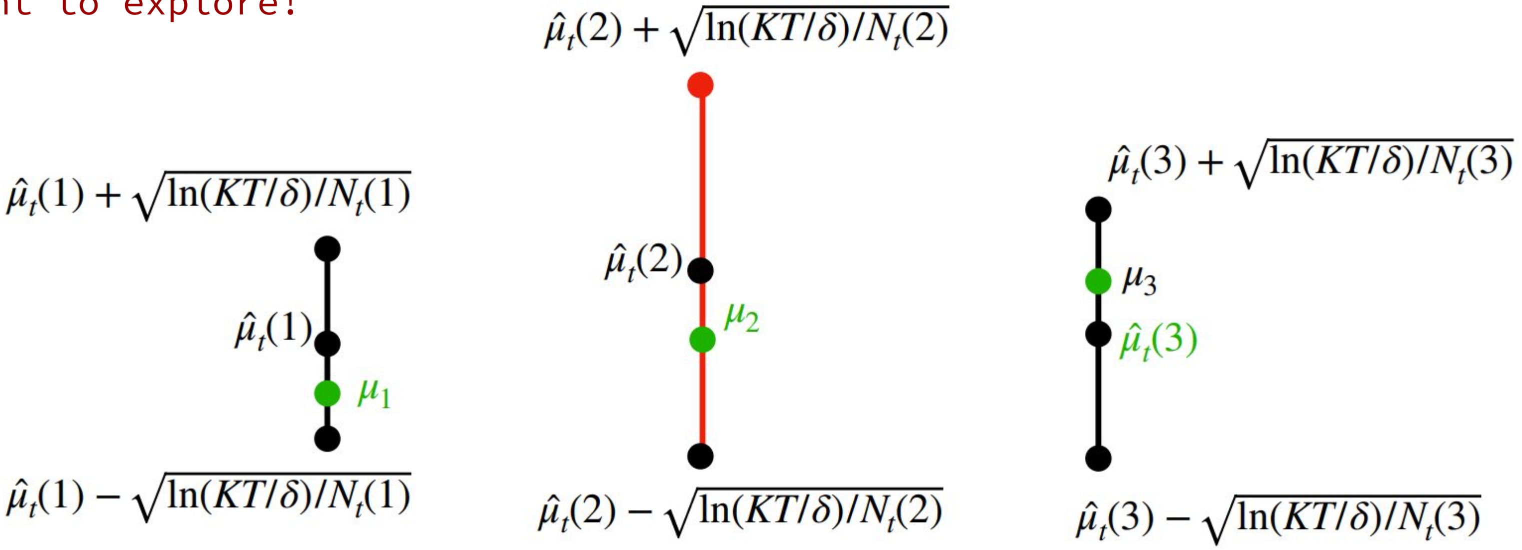
Case 1: large confidence interval, not tried many times (high uncertainty)

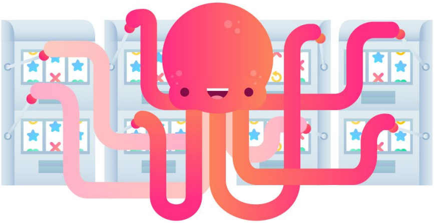


UCB Algorithm: Intuition

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

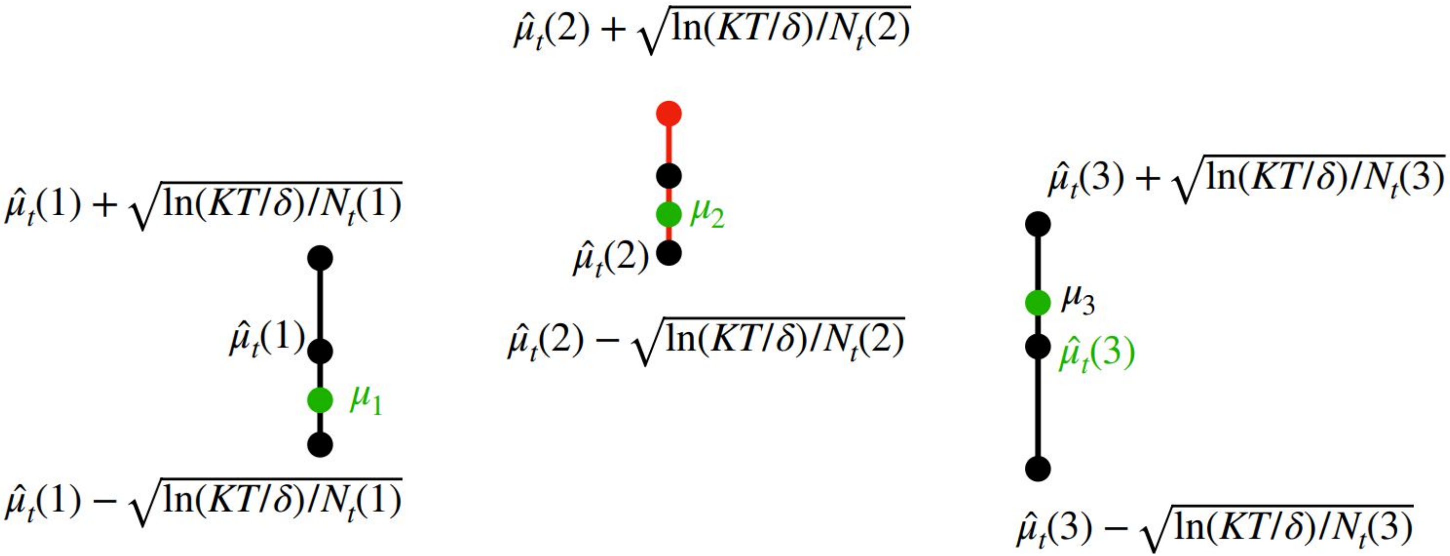
We want to explore!





UCB Algorithm: Intuition

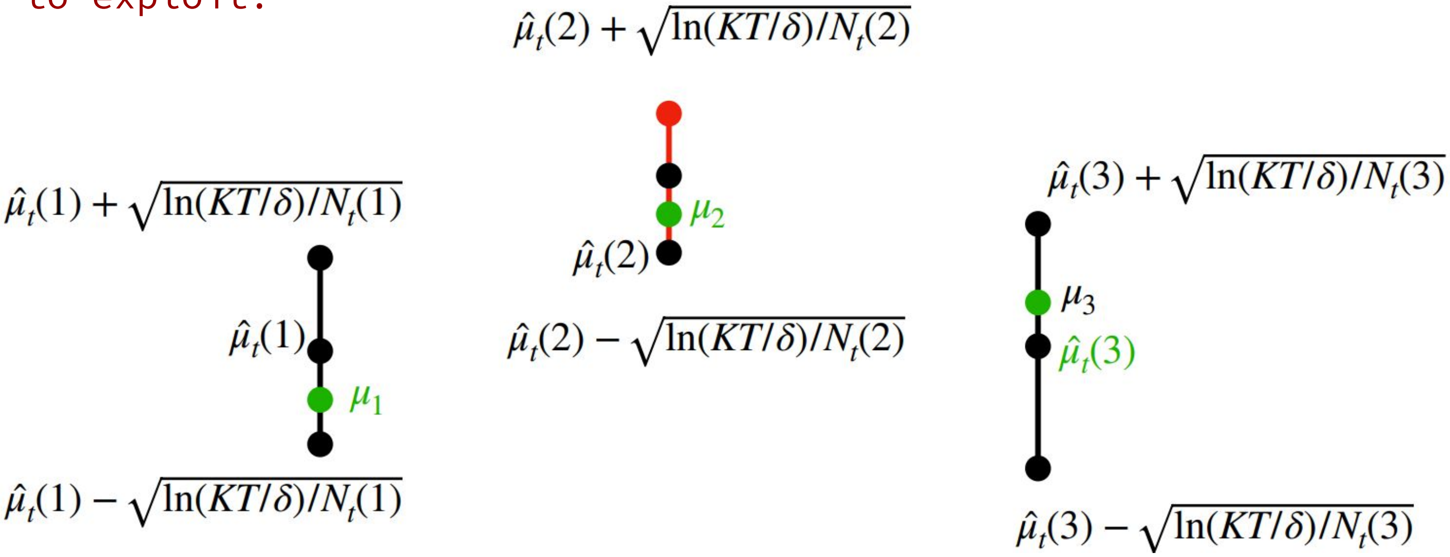
Case 2: small confidence interval, good arm: true mean is high



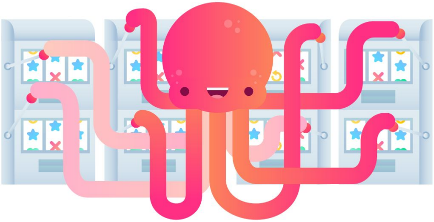
UCB Algorithm: Intuition

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

We want to exploit!



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

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

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

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

UCB Algorithm: Regret-at-t

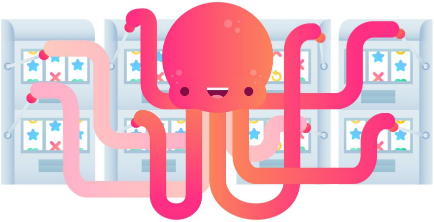
$$I^\star = \arg\max_{i \in [K]} \mu_i$$

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

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

Case 1: 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

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

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

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

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