Exploration in RL: Multi-armed Bandit (Continue)

Recap: MAB

Interactive learning process:

For $t = 0 \rightarrow T - 1$

(# based on historical information)

1. Learner pulls arm $I_t \in \{1,\ldots,K\}$

2. Learner observes an i.i.d reward $r_t \sim \nu_{I_t}$ of arm I_t

unknown

Recap: MAB

Interactive learning process:

For $t = 0 \rightarrow T-1$

(# based on historical information)

1. Learner pulls arm $I_t \in \{1,\ldots,K\}$

2. Learner observes an i.i.d reward $r_t \sim \nu_{I_t}$ of arm I_t

Learning metric:

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

$$\lim_{T\to\infty} \frac{\text{Regret}_T}{T} = 0$$

Recap: MAB

The Explore and Commit Algorithm:

Recap: MAB

The Explore and Commit Algorithm:

For $k = 1 \to K$: (# Exploration phase)

Pull arm-k N times, observe $\{r_i\}_{i=1}^N \sim \nu_k$

Calculate arm k's empirical mean: $\hat{\mu}_k = \sum_{i=1}^N r_i / N$

$\hat{\mu}_k \to \mu_k$

as $N \to \infty$

Recap: MAB

The Explore and Commit Algorithm:

For $k = 1 \to K$: (# Exploration phase)

Pull arm-k N times, observe $\{r_i\}_{i=1}^{N} \sim \nu_k$

Calculate arm k's empirical mean: $\hat{\mu}_k = \sum_{i=1}^{N} r_i / N$

For $t = NK \to T - 1$: (# Exploitation phase)

Pull the best empirical arm, i.e., $I_t = \arg\max_{i \in [K]} \hat{\mu}_i$

Recap: MAB

[Theorem] Fix $\delta \in (0,1)$, set $N = \left(\frac{T\sqrt{\ln(K/\delta)}}{2K}\right)^{2/3}$, with probability at least $1-\delta$, **Explore and Commit** has the following regret:

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

$$\frac{T^{2/3}}{T} = T^{-\frac{1}{3}} \to 0, \quad T \to \infty$$

Question for Today:

$T^{2/3}$

Can we design an algorithm that achieves $\widetilde{O}(\sqrt{T})$ regret?

optimal Regret

Outline:

1. The upper Confidence Bound Algorithm

UCB

2. Analysis of UCB algorithm

Statistics that we maintain during learning:

We maintain the following statistics during the learning process:

At the beginning of iteration $t,$ for all $i \in [K]$, # of times we have tried arm $i,$

Statistics that we maintain during learning:

We maintain the following statistics during the learning process:

At the beginning of iteration t, for all $i \in [K]$, # of times we have tried arm i,

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

Statistics that we maintain during learning:

We maintain the following statistics during the learning process:

At the beginning of iteration t, for all $i \in [K]$, # of times we have tried arm i,

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

and its empirical mean $\hat{\mu}_t(i)$ so far;

Statistics that we maintain during learning:

We maintain the following statistics during the learning process:

At the beginning of iteration t, for all $i \in [K]$, # of times we have tried arm i,

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

$\tau=0$ $\quad$ $t-1$

1 1 2 1 3 3 1

r_0 r_1 r_2 r_3 - - -

and its empirical mean $\hat{\mu}_t(i)$ so far;

$$\text{i.e., } \hat{\mu}_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}\{I_\tau = i\} r_\tau / N_t(i)$$

Recall the Tool for Building Confidence Interval:

[Hoeffding]Given a distribution $\mu \in \Delta([0,1])$, and N i.i.d samples $\{r_i\}_{i=1}^{N} \sim \mu$, w/ probability at least $1-\delta$, we have:

$$\left| \underbrace{\sum_{i=1}^{N} r_i/N}_{\text{estimated mean}} - \mu \right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$

Recall the Tool for Building Confidence Interval:

[Hoeffding]Given a distribution $\mu \in \Delta([0,1])$, and N i.i.d samples $\{r_i\}_{i=1}^N \sim \mu$, w/ probability at least $1-\delta$, we have:

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

Thus, we know that for all iteration t, we have the for all $i \in [K]$, w/ prob $1-\delta$,

$$|\underbrace{\hat{\mu}_t(i)}_{\text{estimated mean}} - \underbrace{\mu_i}_{\text{GT}}| \leq \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$$

Recall the Tool for Building Confidence Interval:

[Hoeffding]Given a distribution $\mu \in \Delta([0,1])$, and N i.i.d samples $\{r_i\}_{i=1}^N \sim \mu$, w/ probability at least $1-\delta$, we have:

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

Thus, we know that for all iteration t, we have the for all $i \in [K]$, w/ prob $1-\delta$,

$$|\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$$

(Note that I applied union bound over all $t \in [T]$ and all $i \in [K]$, but let's not worry too much about $\log$ terms—*details are in reading material in case you are interested*)

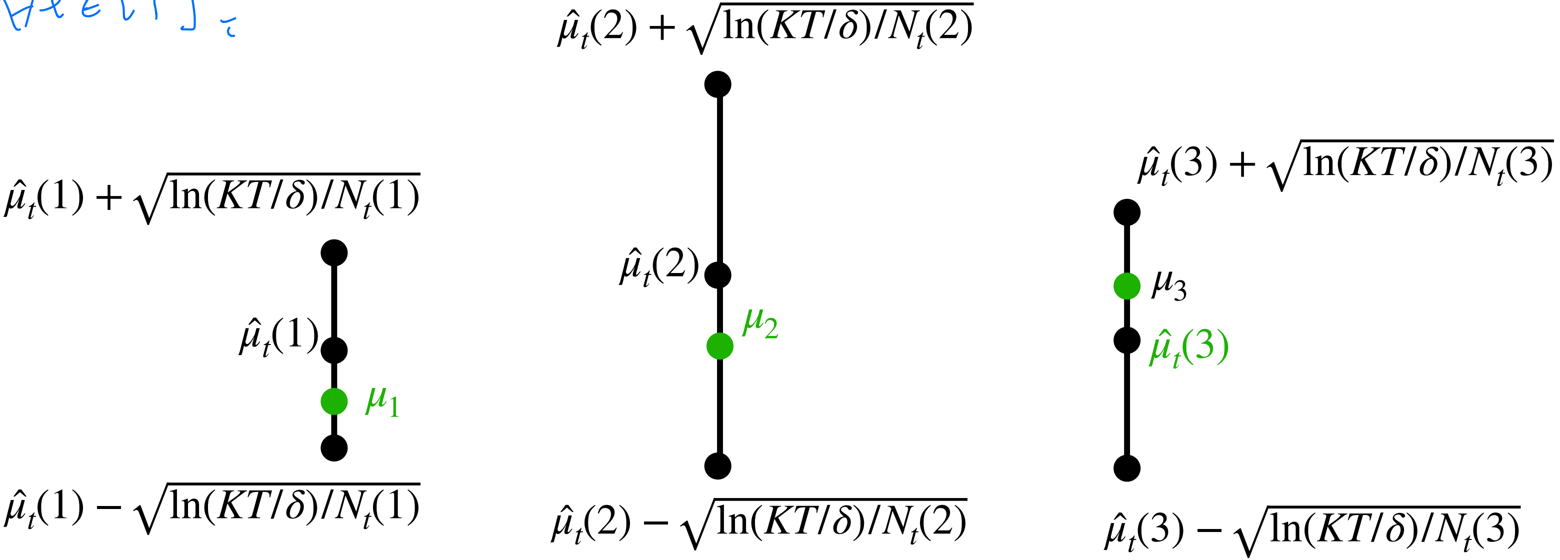
Summary so far:

W/ high prob, we have valid confidence intervals for all iteration t, and all arm i:

Summary so far:

W/ high prob, we have valid confidence intervals for all iteration t, and all arm i:

$\forall t \in [T]$ i

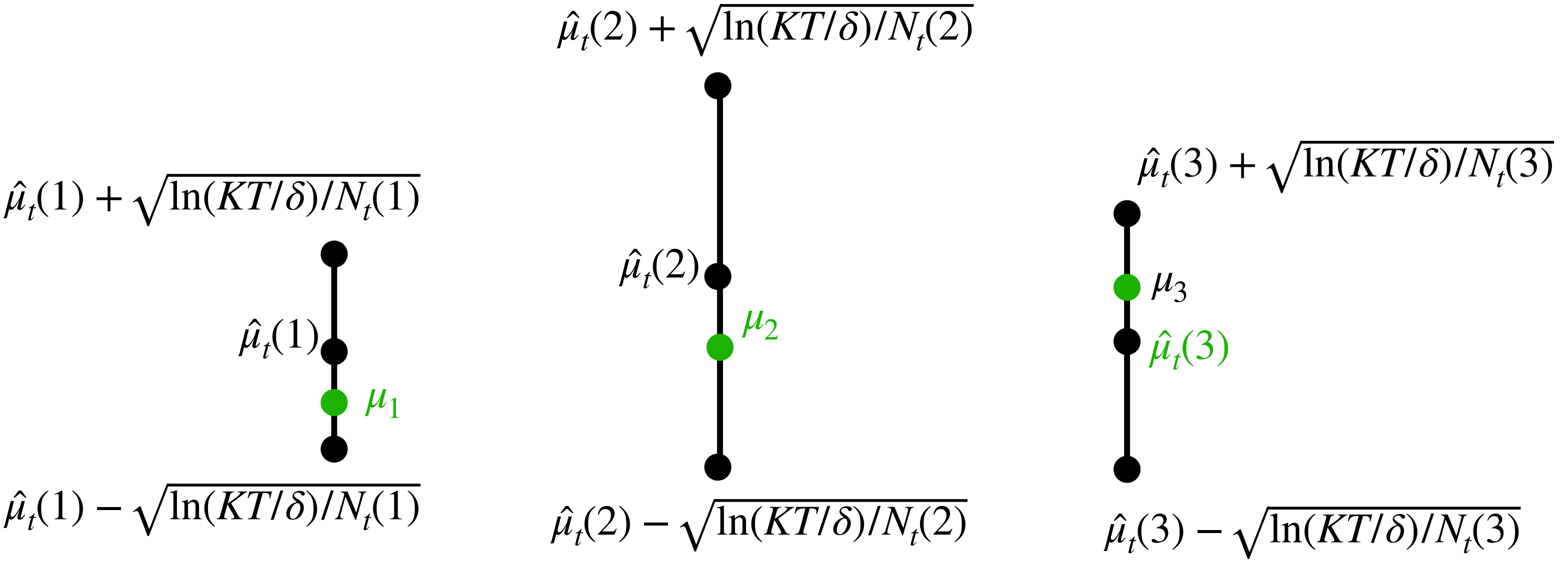


UCB: Optimism in the face of Uncertainty

Given the confidence interval, we pick arm that has the **highest Upper-Conf-Bound:**

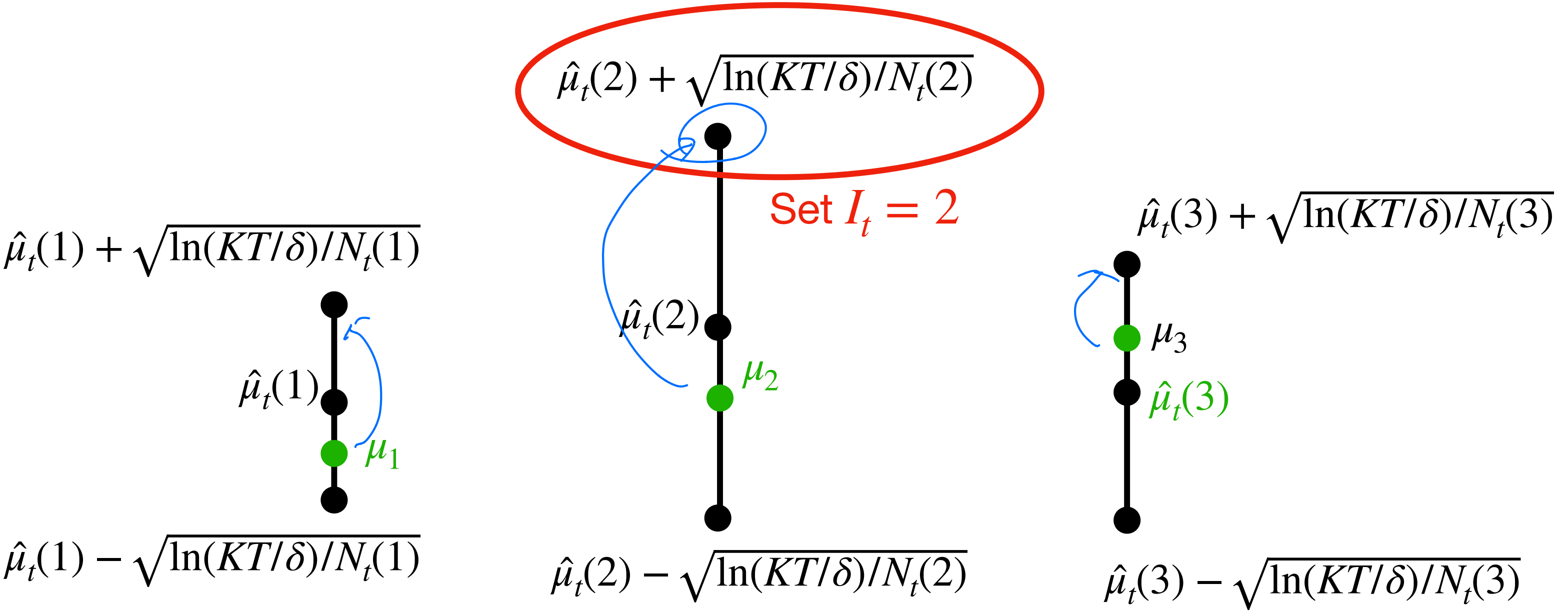
UCB: Optimism in the face of Uncertainty

Given the confidence interval, we pick arm that has the **highest Upper-Conf-Bound:**



UCB: Optimism in the face of Uncertainty

Given the confidence interval, we pick arm that has the **highest Upper-Conf-Bound:**



Put things together: UCB Algorithm:

For $t = 0 \rightarrow T - 1$:

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

Put things together: UCB Algorithm:

For $t = 0 \rightarrow T-1$: estimated mean   half conf-Interval

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

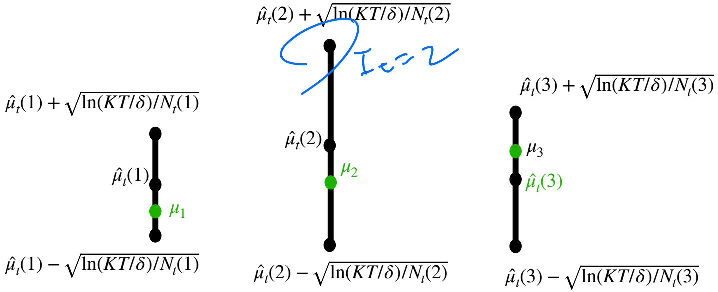
(# Upper-conf-bound of arm i)

Put things together: UCB Algorithm:

For $t = 0 \rightarrow T - 1$:

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

(# Upper-conf-bound of arm i)



Put things together: UCB Algorithm:

$T > K$

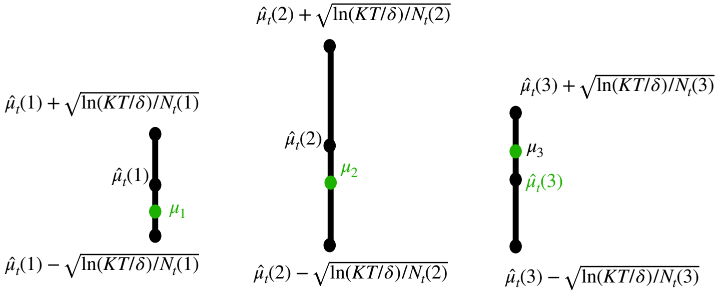
Warmstart: For the first K iterations let's pull each arm once !!

↳ Total Regret $= K-1$

For $t = 0 \to T-1$:

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

(# Upper-conf-bound of arm i)



"Reward Bonus": $\sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Bonus ↑, if $N_t(i)$ ↓

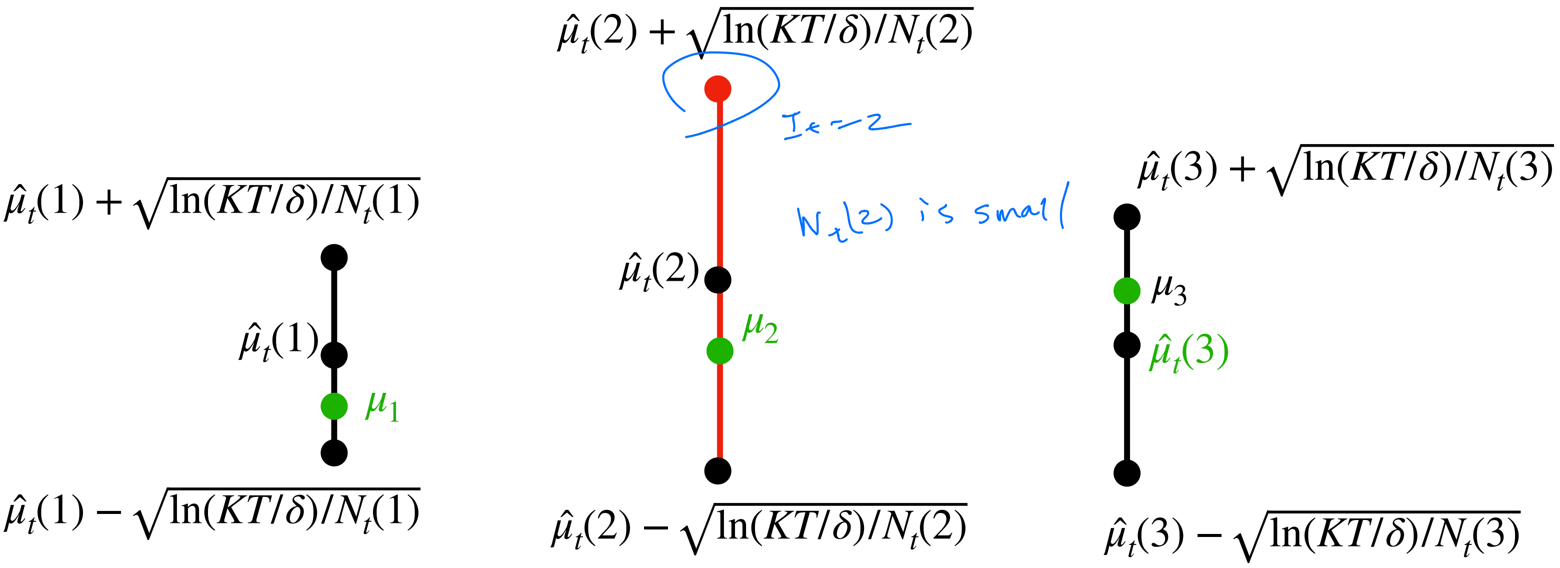
Outline:

✓ 1. The upper Confidence Bound Algorithm

2. Analysis of UCB algorithm

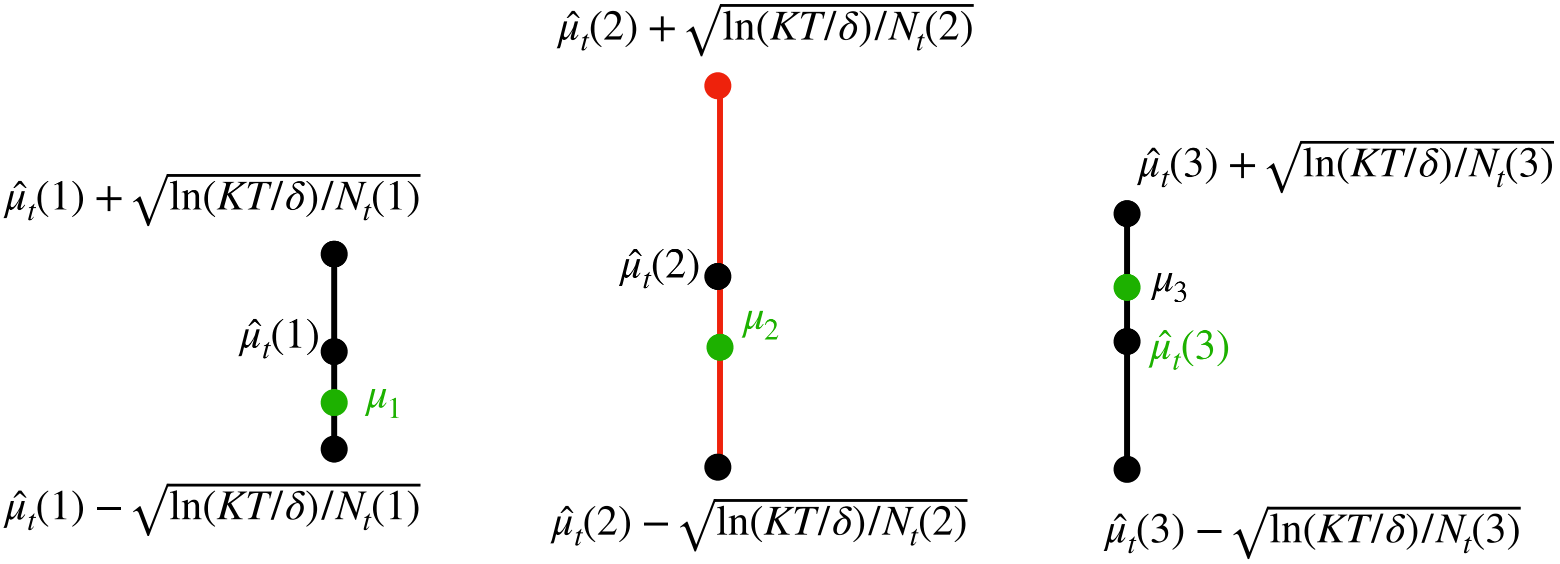
Intuitive Explanation of UCB

Intuitive Explanation of UCB



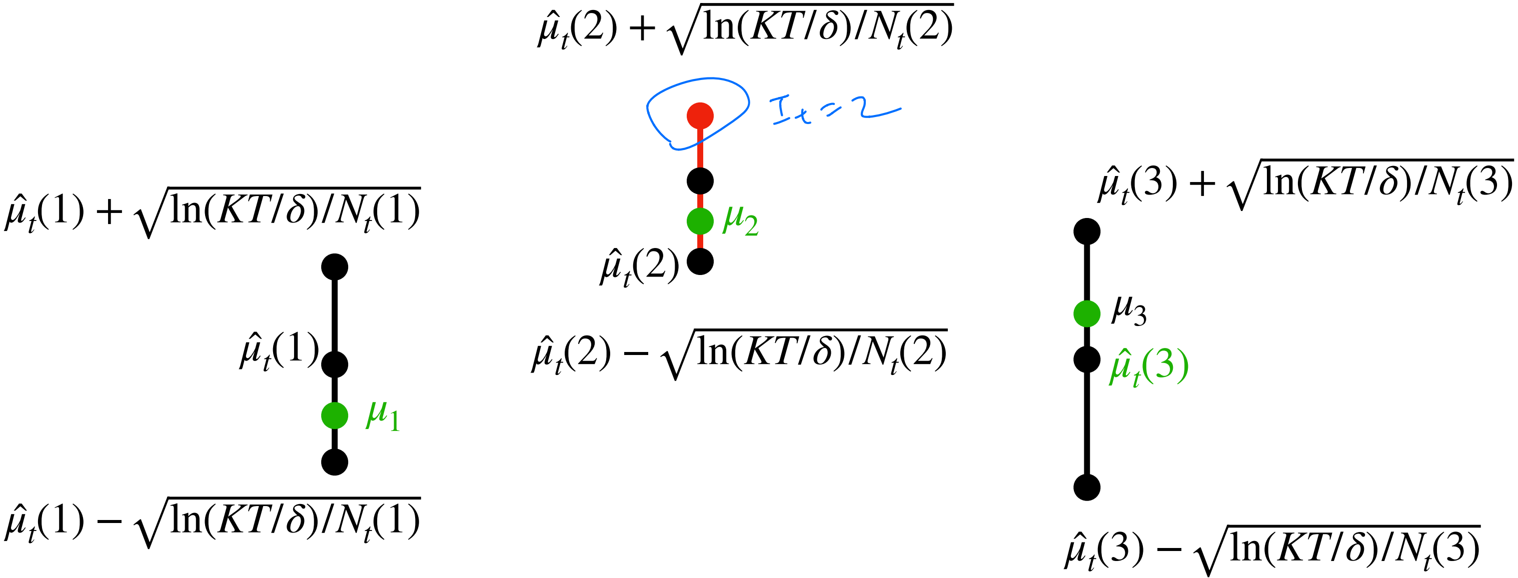
Intuitive Explanation of UCB

Case 1: it has large conf-interval, which means that it has not been tried many times yet (high uncertainty)



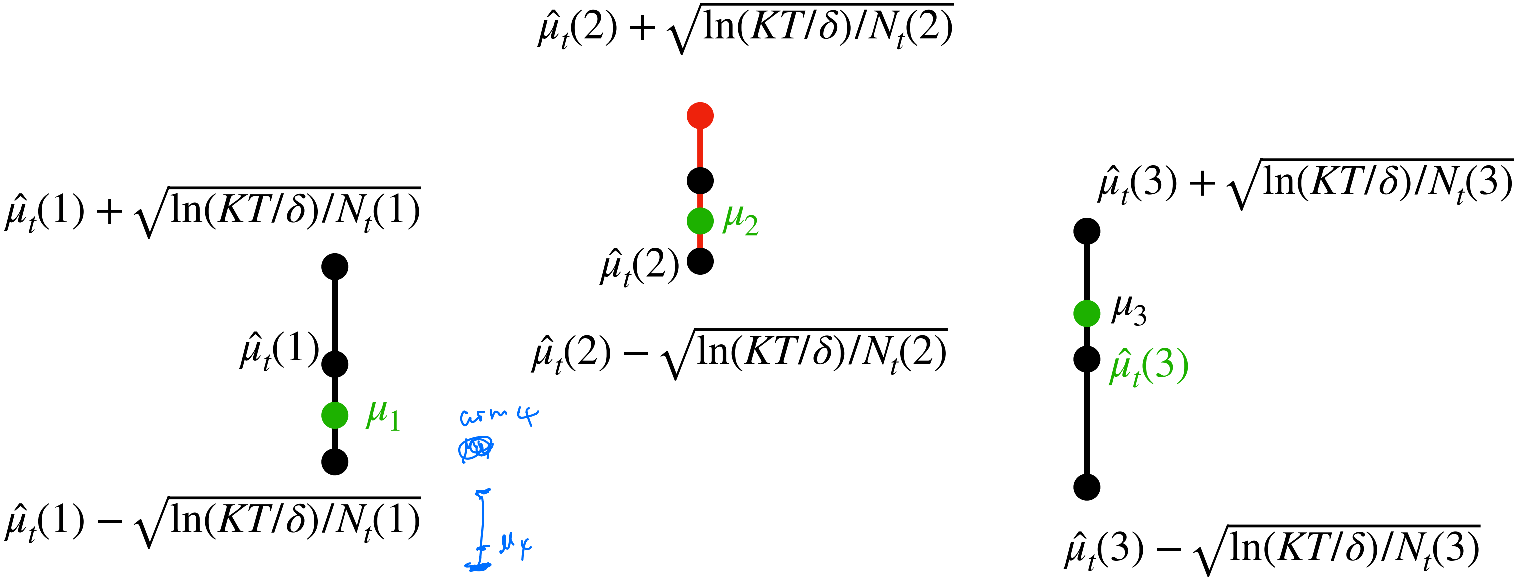
Intuitive Explanation of UCB

Intuitive Explanation of UCB



Intuitive Explanation of UCB

Case 2: it has low uncertainty, then it is simply a good arm, i.e., it's true mean is high!



Explore and Exploration Tradeoff

Case 1: I_t has large conf-interval, which means that it has not been tried many times yet (high uncertainty)

Thus, we do exploration in this case!

$\ln(1/\delta)$

Explore and Exploration Tradeoff

Case 1: I_t has large conf-interval, which means that it has not been tried many times yet (high uncertainty)

Thus, we do exploration in this case!

Case 2: I_t has small conf-interval, then it is simply a good arm, i.e., it’s true mean is pretty high!

Thus, we do exploitation in this case!

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \underbrace{\hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}}_{\mathrm{UCB}[i]}$

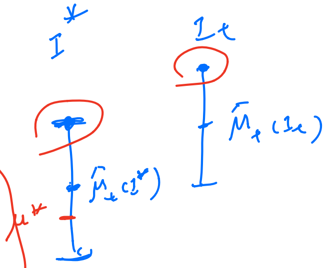
Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i\in[K]} \mu_i$; recall $I_t = \arg\max_{i\in[K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-t $= \mu^\star - \mu_{I_t}$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\dfrac{\ln(KT/\delta)}{N_t(i)}}$



Regret-at-t $= \mu^\star - \mu_{I_t}$

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$\mathrm{UCB}[I_t] \geq \mathrm{UCB}[I^*] \geq \mu^*$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-t $= \mu^\star - \mu_{I_t}$

Q: why?

$$\le \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i\in[K]} \mu_i$; recall $I_t = \arg\max_{i\in[K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

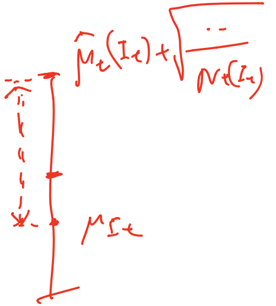
Regret-at-t $= \mu^\star - \mu_{I_t}$

Q: why?

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\mu_{I_t} \geq \underbrace{\hat{\mu}_t(I_t) - \sqrt{\frac{\ln(\ldots)}{N_t(I_t)}}}_{LCB}$$

$$\leq \underbrace{2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}}_{\text{Length-of-Conf-Interval of } I_t \checkmark}$$



Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i\in[K]} \mu_i$; recall $I_t = \arg\max_{i\in[K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-t $= \mu^\star - \mu_{I_t}$

Q: why?

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\leq \underbrace{2\sqrt{\frac{\ln(TK/\delta)}{\underbrace{N_t(I_t)}_{\text{small}}}}}_{\text{big}}$$

Case 1: $N_t(I_t)$ is small

(i.e., uncertainty about I_t is large);

We pay regret, BUT we **explore** here,

as we just tried I_t at iter t!

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

$0 \le \mu^\star - \mu_{I_t} \le 2\sqrt{\frac{\ln(\cdots)}{N_t(I_t)}}$

$|\mu^\star - \mu_{I_t}| \le 2\sqrt{\frac{\ln(\cdots)}{N_t(I_t)}}$ ↓ small

Regret-at-t $= \mu^\star - \mu_{I_t}$

$$\le \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\le 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$ ↑ big ↓ small

Case 2: $N_t(I_t)$ is large, i.e., conf-interval of I_t is small,

Then we **exploit** here, as I_t is pretty good (the gap between $\mu^\star$ & μ_{I_t} is small)!

Let's formalize the intuition

Finally, let's add all per-iter regret together:

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

$$\leq \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\overset{=}{\cancel{\leq}}\ 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}$$

Let's formalize the intuition

Finally, let's add all per-iter regret together:

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

$$\leq \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\leq 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}$$

Lemma (optional):

$$\sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}} \leq O\left(\sqrt{KT}\right)$$

UCB Regret:

[Theorem (informal)] With high probability, UCB has the following regret:

$$\text{Regret}_T = \widetilde{O}\left(\underbrace{\sqrt{KT}}_{\text{optimal}}\right)$$

(See reading material for more details)

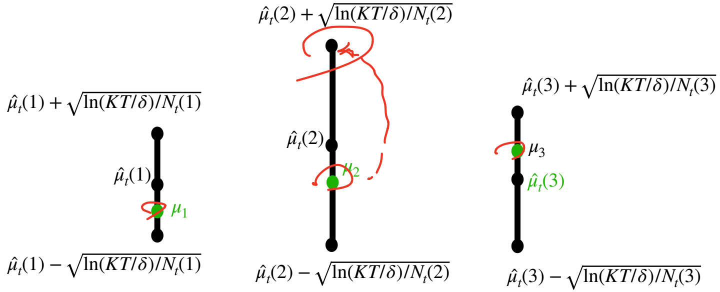
Summary for Today:

UCB algorithm: ***Principle of Optimism in the face of Uncertainty***

For $t = 0 \rightarrow T - 1$:

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

(# Upper-conf-bound of arm i)



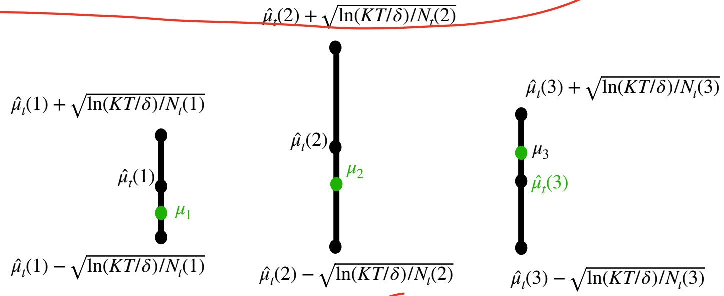
Summary for Today:

UCB algorithm: *Principle of Optimism in the face of Uncertainty*

For $t = 0 \rightarrow T - 1$:

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

(# Upper-conf-bound of arm i)



Analysis Intuition:

Case 1: the arm I_t has high uncertainty (we explore)

Case 2: the arm I_t has low uncertainty, then it must be a near-optimal arm (i.e.,

$|\mu_{I_t} - \mu_{I^*}| \approx$ small| exploit)

$X_t \qquad \pi^*(X_t) \to I_t \quad I_t \in \{1, \ldots K\}$

$A \qquad A$

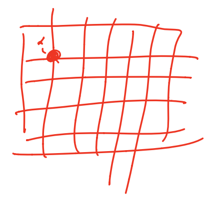
$b(s,a) \to [0,1]$ $\qquad \sqrt{\frac{1}{N(s,a)}}$

$b(s,a)$ is big if (s,a) is under explored

$b(s,a)$ is small if (s,a) is explored many times

$$\checkmark \; \mathbb{E}_{s \sim d_\mu^\pi} \left[\nabla_\theta \ln \pi_\theta(a|s) \cdot \left[\sum_{h=0}^{H-1} r(s_h, a_h) + \boxed{\lambda \cdot b(s_h, a_h)} \right] \right]$$

discrete mountain car



$\frac{1}{N(s,a)}$