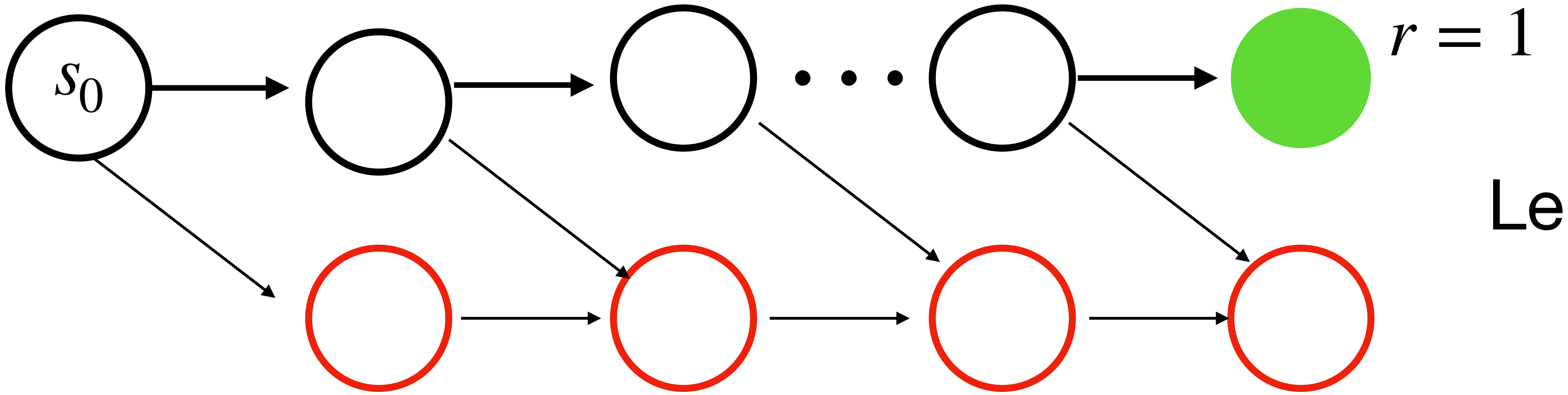
Multi-armed Bandits

CS 6789: Foundations of Reinforcement Learning

The need for Exploration in RL:

The Combination Lock Example (i.e., the sparse reward problem)

(1) We have reward zero everywhere except at the goal (the right end);
(2) Every black node, one of the two actions will lead the agent to the dead state (red)

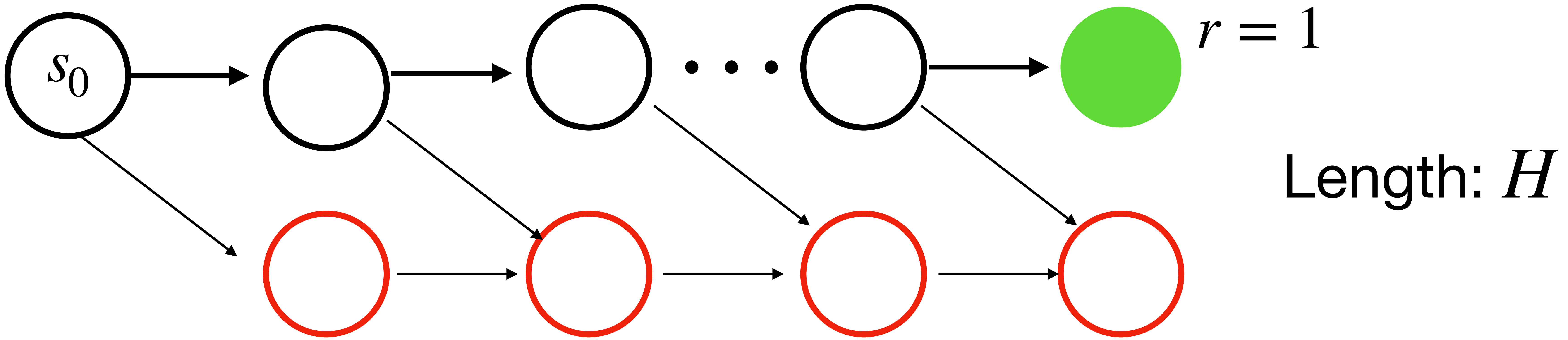


Length: H

The need for Exploration in RL:

The Combination Lock Example (i.e., the sparse reward problem)

(1) We have reward zero everywhere except at the goal (the right end);
(2) Every black node, one of the two actions will lead the agent to the dead state (red)



What is the probability of a random policy generating a trajectory that hits the goal?

Exploration!

We need to perform systematic exploration,
i.e., remember where we visited, and purposely try to visit unexplored regions..

What we will do today:

Study Exploration in a very simple MDP:

$$\mathcal{M} = \{s_0, \{a_1, \ldots, a_K\}, H = 1, R\}$$

i.e., MDP with one state, one-step transition, and K actions

This is also called Multi-armed Bandits

Plan for today:

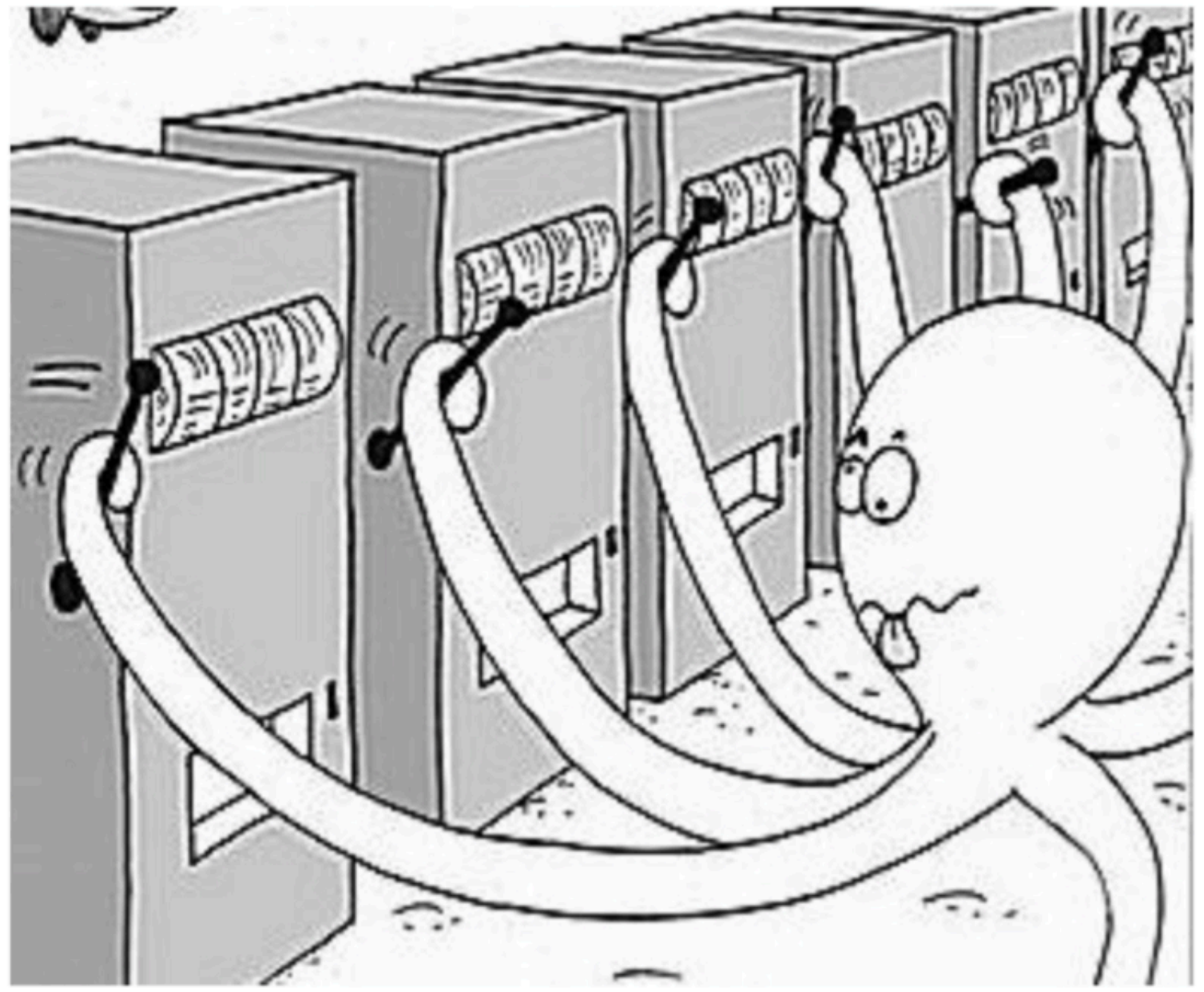
1. Introduction of MAB

2. Attempt 1: Greedy Algorithm (a bad algorithm)

3. Attempt 2: Explore and Commit

4. Attempt 3: Upper Confidence Bound (UCB) Algorithm

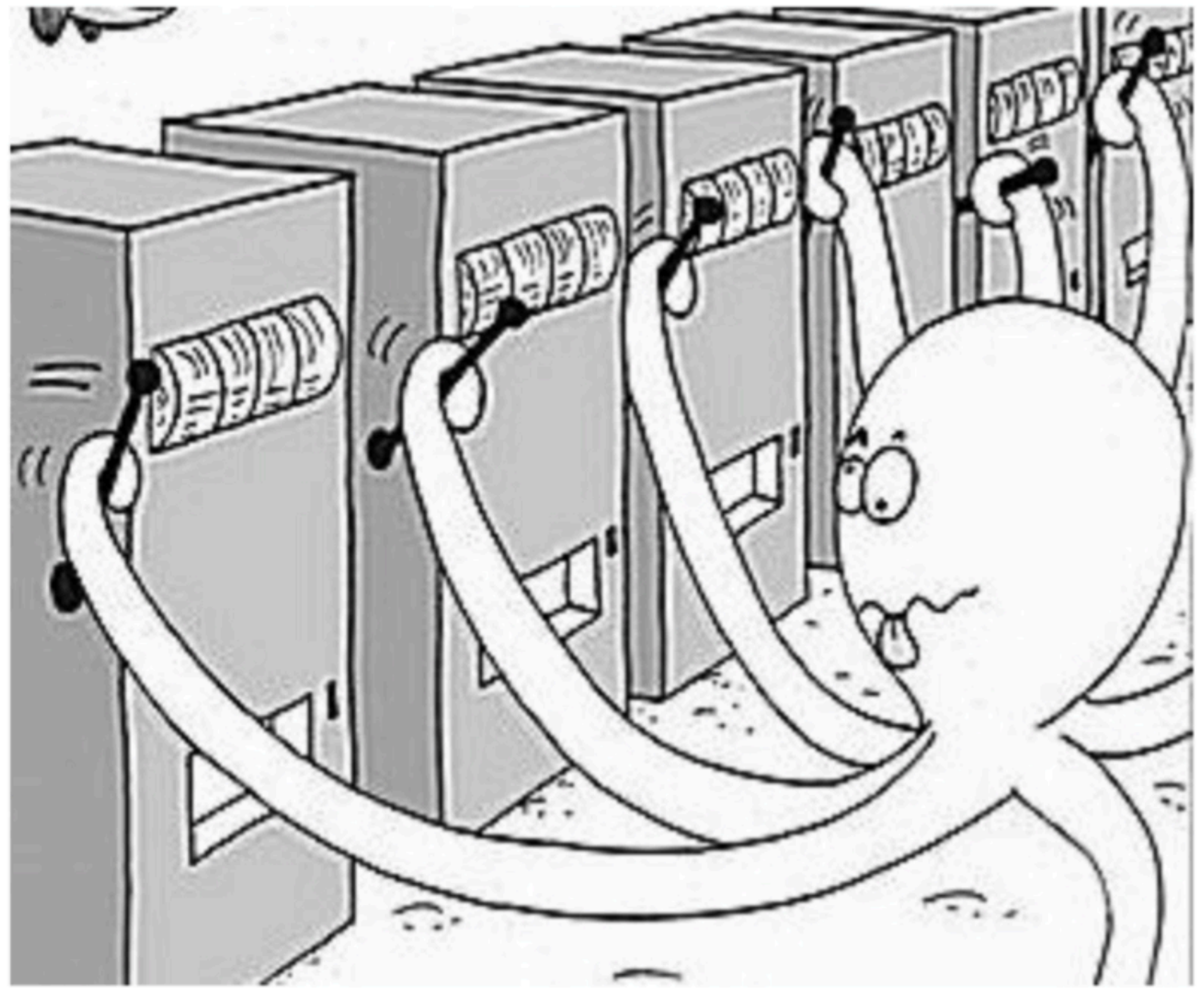
Intro to MAB



Setting:

We have K many arms: $a_1, \ldots, a_K$

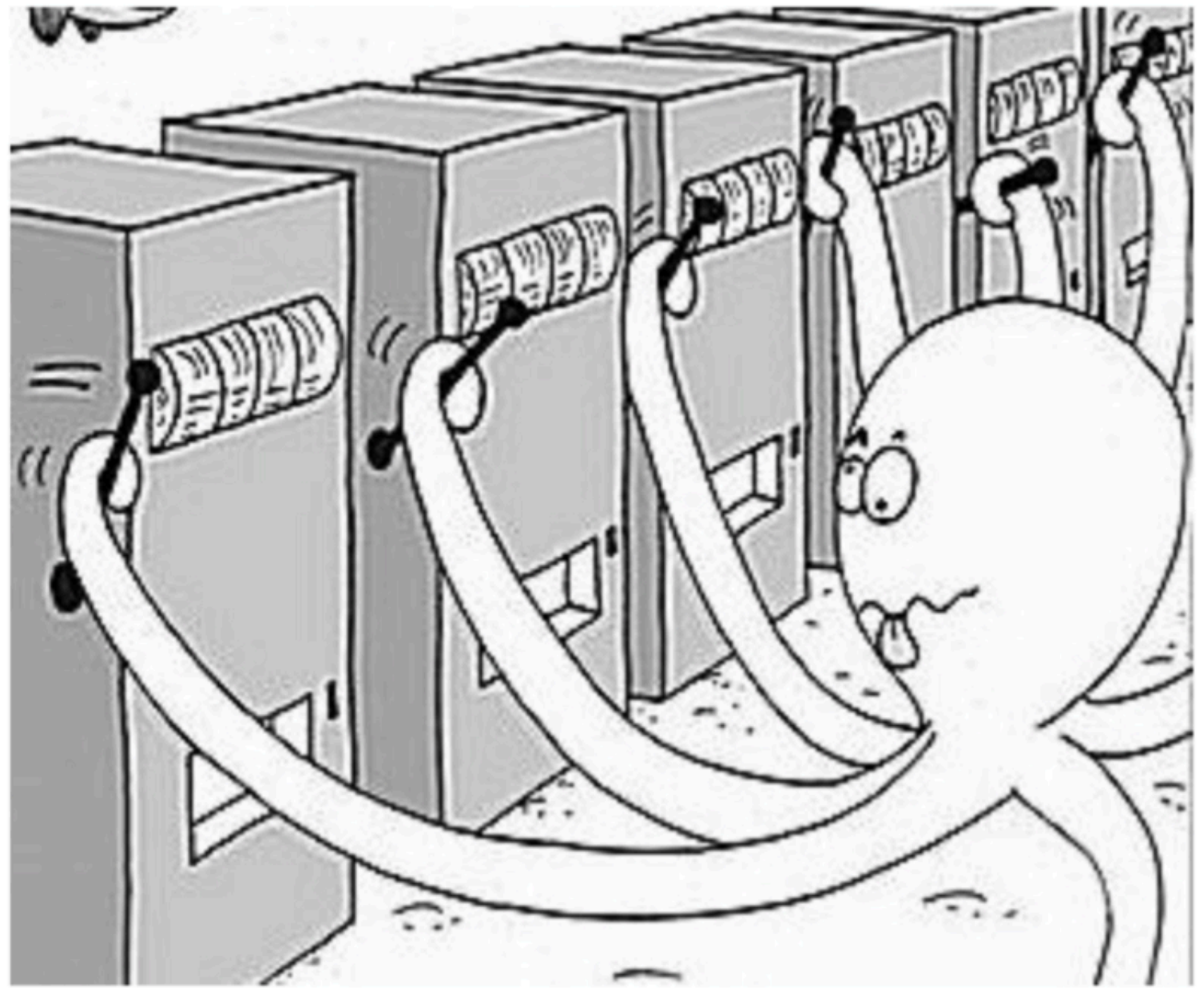
Intro to MAB



Setting:

We have K many arms: $a_1, \ldots, a_K$

Each arm has a unknown reward distribution, i.e., $\nu_i \in \Delta([0,1])$,
w/ mean $\mu_i = \mathbb{E}_{r \sim \nu_i}[r]$



Intro to MAB

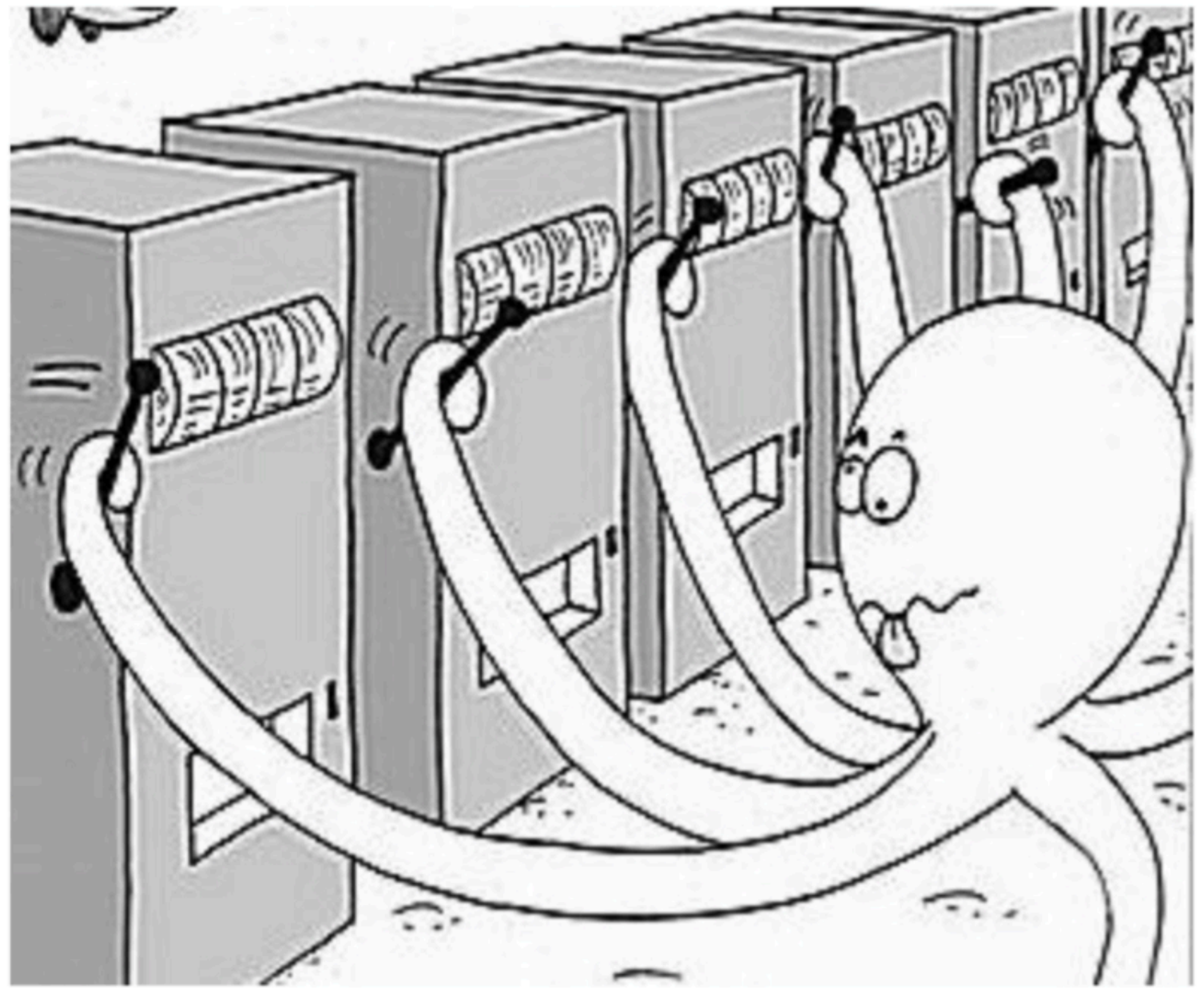
Setting:

We have K many arms: $a_1, \ldots, a_K$

Each arm has a unknown reward distribution, i.e., $\nu_i \in \Delta([0,1])$,
w/ mean $\mu_i = \mathbb{E}_{r\sim\nu_i}[r]$

Example: a_i has a Bernoulli distribution ν_i w/ mean $\mu_i := p$:

Intro to MAB



Setting:

We have K many arms: $a_1, \ldots, a_K$

Each arm has a unknown reward distribution, i.e., $\nu_i \in \Delta([0,1])$,
w/ mean $\mu_i = \mathbb{E}_{r \sim \nu_i}[r]$

Example: a_i has a Bernoulli distribution ν_i w/ mean $\mu_i := p$:

Every time we pull arm a_i, we observe an i.i.d reward $r = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1 - p \end{cases}$

Intro to MAB

Applications on online advertisement:

Arms correspond to Ads

Each arm has **click-through-rate** (CTR): probability of getting clicked (unknown)

Intro to MAB

Applications on online advertisement:

Arms correspond to Ads

Each arm has **click-through-rate** (CTR): probability of getting clicked (unknown)

A learning system aims to maximize CTR in a long run:

Intro to MAB

Applications on online advertisement:

Arms correspond to Ads

Each arm has **click-through-rate** (CTR): probability of getting clicked (unknown)

A learning system aims to maximize CTR in a long run:

1. **Try** an Ad (pull an arm)

Intro to MAB

Applications on online advertisement:

Arms correspond to Ads

Each arm has **click-through-rate** (CTR): probability of getting clicked (unknown)

A learning system aims to maximize CTR in a long run:

1. **Try** an Ad (pull an arm)
2. **Observe** if it is clicked (see a zero-one **reward**)

Intro to MAB

Applications on online advertisement:

Arms correspond to Ads

Each arm has **click-through-rate** (CTR): probability of getting clicked (unknown)

A learning system aims to maximize CTR in a long run:

1. **Try** an Ad (pull an arm)

2. **Observe** if it is clicked (see a zero-one **reward**)

3. **Update**: Decide what ad to recommend for next round

Intro to MAB

More formally, we have the following interactive learning process:

For $t = 0 \rightarrow T - 1$

Intro to MAB

More formally, we have the following interactive learning process:

For $t = 0 \to T - 1$

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

Intro to MAB

More formally, we have the following interactive learning process:

For $t = 0 \rightarrow T - 1$

(# based on historical information)

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

Intro to MAB

More formally, we have the following 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

Intro to MAB

More formally, we have the following 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

Note: each iteration, we do not observe rewards of arms that we did not try

Intro to MAB

More formally, we have the following learning objective:

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

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

Intro to MAB

More formally, we have the following learning objective:

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

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

Total expected reward if we pulled best arm over T rounds

Intro to MAB

More formally, we have the following learning objective:

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

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

Total expected reward if we pulled best arm over T rounds

Total expected reward of the arms we pulled over T rounds

Intro to MAB

More formally, we have the following learning objective:

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

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

Total expected reward if we pulled best arm over T rounds

Total expected reward of the arms we pulled over T rounds

Goal: no-regret, i.e., $\text{Regret}_T / T \to 0$, as $T \to \infty$

Intro to MAB

Why the problem is hard?

Exploration and Exploitation Tradeoff:

Intro to MAB

Why the problem is hard?

Exploration and Exploitation Tradeoff:

Every round, we need to ask ourselves:

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**)?

Plan for today:

1. Introduction of MAB

2. Attempt 1: Greedy Algorithm (a bad algorithm)

3. Attempt 2: Explore and Exploit

4. Attempt 3: Upper Confidence Bound (UCB) Algorithm

Attempt 1: Greedy Algorithm

Alg: try each arm once, and then commit to the one that has the **highest observed** reward

Attempt 1: Greedy Algorithm

Alg: try each arm once, and then commit to the one that has the **highest observed** reward

Q: what could be wrong?

Attempt 1: Greedy Algorithm

Alg: try each arm once, and then commit to the one that has the **highest observed** reward

Q: what could be wrong?

A bad arm (i.e., low μ_i) may generate a high reward by chance!

(recall we have $r \sim \nu$, i.i.d)

Attempt 1: Greedy Algorithm

More concretely, let's say we have two arms a_1, a_2:

Reward dist for a_1: w/ prob 60%, $r = 1$; else $r = 0$

Reward dist for a_2: w/ prob 40%, $r = 1$; else $r = 0$

Attempt 1: Greedy Algorithm

More concretely, let's say we have two arms a_1, a_2:

Reward dist for a_1: w/ prob 60%, $r = 1$; else $r = 0$

Reward dist for a_2: w/ prob 40%, $r = 1$; else $r = 0$

Clearly a_1 is a better arm!

Attempt 1: Greedy Algorithm

More concretely, let's say we have two arms a_1, a_2:

Reward dist for a_1: w/ prob 60%, $r = 1$; else $r = 0$

Reward dist for a_2: w/ prob 40%, $r = 1$; else $r = 0$

Clearly a_1 is a better arm!

But try a_1, a_2 once, with probability 16%, we will observe reward pair $(0,1)$

Attempt 1: Greedy Algorithm

More concretely, let's say we have two arms a_1, a_2:

Reward dist for a_1: w/ prob 60%, $r = 1$; else $r = 0$

Reward dist for a_2: w/ prob 40%, $r = 1$; else $r = 0$

Clearly a_1 is a better arm!

But try a_1, a_2 once, with probability 16%, we will observe reward pair $(0,1)$

The greedy alg will pick a_2—**loosing expected reward 0.2 every time in the future**

Plan for today:

1. Introduction of MAB

2. Attempt 1: Greedy Algorithm
(a bad algorithm: constant regret)

3. Attempt 2: Explore and Commit

4. Attempt 3: Upper Confidence Bound (UCB) Algorithm

What lessons we learned from the Greedy Alg:

Due to randomness in the reward distribution, trying each arm once is not enough, i.e., observed single reward may be far away from the mean

What lessons we learned from the Greedy Alg:

Due to randomness in the reward distribution, trying each arm once is not enough, i.e., observed single reward may be far away from the mean

Q: what's the fix here?

What lessons we learned from the Greedy Alg:

Due to randomness in the reward distribution, trying each arm once is not enough, i.e., observed single reward may be far away from the mean

Q: what's the fix here?

Yes, let's (1) try each arm multiple times, (2) compute the empirical mean of each arm, (3) commit to the one that has the highest empirical mean

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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

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

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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

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

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

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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)

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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

Alg: Explore and Commit:

Algorithm hyper parameter $N < T/K$ (we assume $T >> K$)

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

Q: how to set N?

Statistical Tools:

1. Hoeffding inequality (optional, no need to remember or understand it)

Statistical Tools:

1. Hoeffding inequality (optional, no need to remember or understand it)

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:

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

Statistical Tools:

1. Hoeffding inequality (optional, no need to remember or understand it)

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:

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

i.e., this gives us a confidence interval:

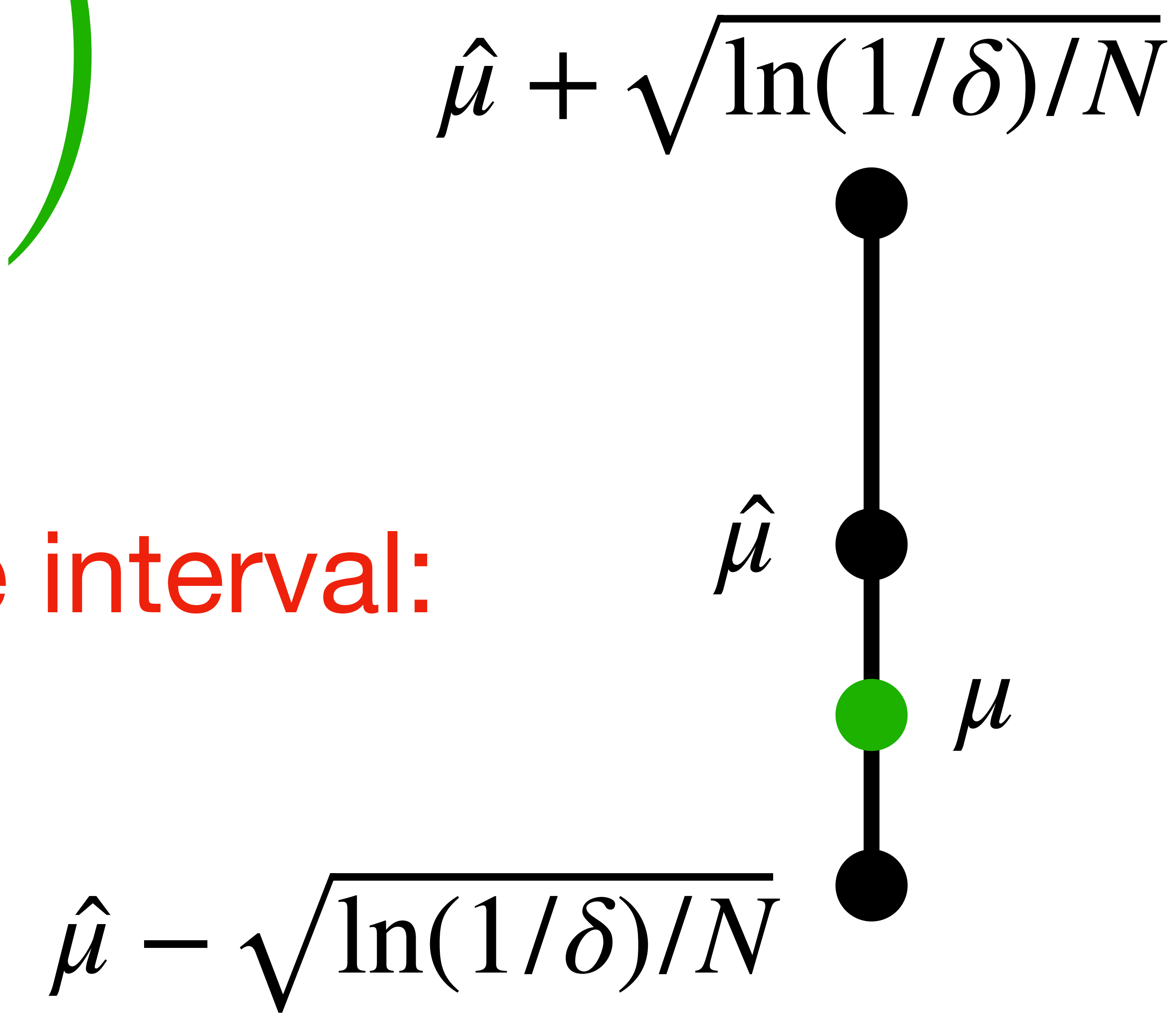
Statistical Tools:

1. Hoeffding inequality (optional, no need to remember or understand it)

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:

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

i.e., this gives us a confidence interval:



Statistical Tools:

Statistical Tools:

Combine Hoeffding and Union Bound, we have:

Statistical Tools:

Combine Hoeffding and Union Bound, we have:

After the Exploration phase, with probability at least 1-δ, **for all arm** $k \in [K]$, we have:

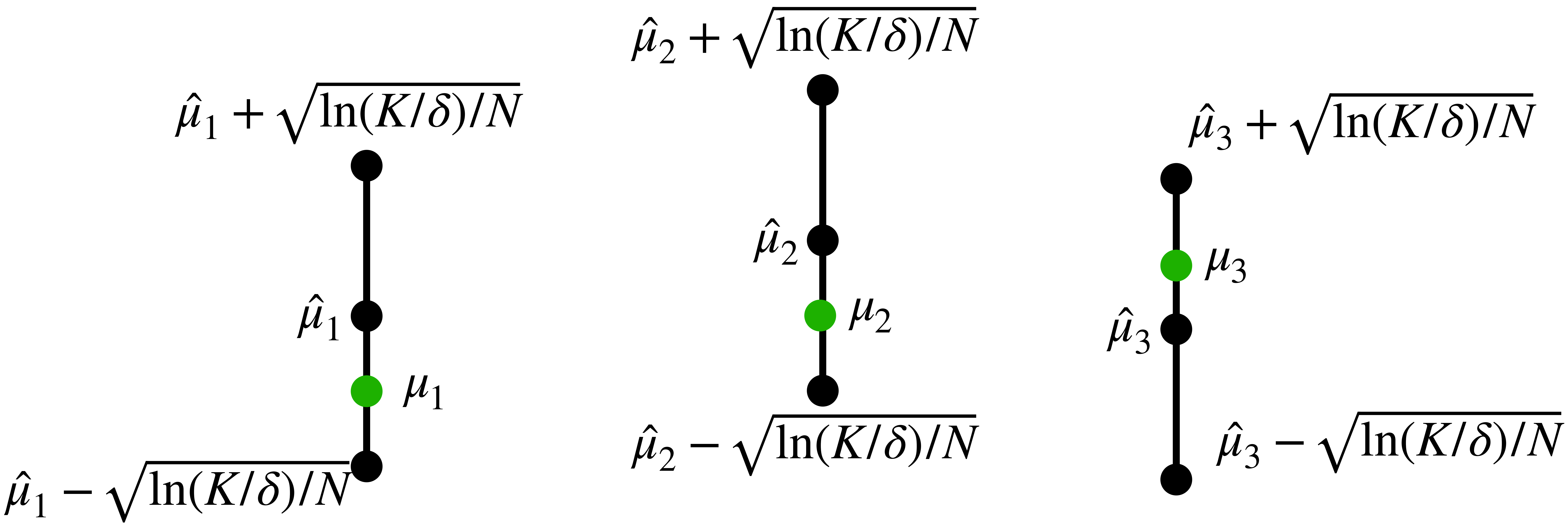
$$\left|\hat{\mu}_k - \mu_k\right| \leq O\left(\sqrt{\frac{\ln(K/\delta)}{N}}\right)$$

Statistical Tools:

Combine Hoeffding and Union Bound, we have:

After the Exploration phase, with probability at least 1-δ, **for all arm** $k \in [K]$, we have:

$$\left|\hat{\mu}_k - \mu_k\right| \leq O\left(\sqrt{\frac{\ln(K/\delta)}{N}}\right)$$



Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

1. What's the worst possible regret in the exploration phase:

Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

1. What's the worst possible regret in the exploration phase:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

1. What's the worst possible regret in the exploration phase:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

2. What's the regret in the exploitation phase:

Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

1. What's the worst possible regret in the exploration phase:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

2. What's the regret in the exploitation phase:

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

Calculate the final regret:

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

1. What's the worst possible regret in the exploration phase:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

2. What's the regret in the exploitation phase:

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

Let's now bound $\text{Regret}_{exploit}$

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

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

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

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

$$= \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + 2\sqrt{\ln(K/\delta)/N}$$

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

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

$$= \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + 2\sqrt{\ln(K/\delta)/N}$$

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

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

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

$$= \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + 2\sqrt{\ln(K/\delta)/N}$$

Q: why?

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

Calculate the regret in the exploitation phase

Denote empirical best arm $\hat{I} = \arg\max_{i \in [K]} \hat{\mu}_i$, and THE best arm $I^\star = \arg\max_{i \in [K]} \mu_i$

What's the regret in the exploitation phase:

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

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

$$= \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + 2\sqrt{\ln(K/\delta)/N}$$

Q: why?

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

$$\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right) \leq 2T\sqrt{\frac{\ln(K/\delta)}{N}}$$

Finally, combine two regret together:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

$$\text{Regret}_{exploit} \leq (T-NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right) \leq T\sqrt{\frac{\ln(K/\delta)}{N}}$$

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

Finally, combine two regret together:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

$$\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right) \leq T\sqrt{\frac{\ln(K/\delta)}{N}}$$

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

Minimize the upper bound via optimizing N:

Finally, combine two regret together:

$$\text{Regret}_{explore} \leq N(K-1) \leq NK$$

$$\text{Regret}_{exploit} \leq (T - NK)\left(\mu_{I^\star} - \mu_{\hat{I}}\right) \leq T\sqrt{\frac{\ln(K/\delta)}{N}}$$

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

Minimize the upper bound via optimizing N:

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

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

To conclude on Explore then Commit:

[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:

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

Q: can we do better, particularly, can we get $\sqrt{T}$ regret bound?

Plan for today:

1. Introduction of MAB

2. Attempt 1: Greedy Algorithm
(a bad algorithm: constant regret)

3. Attempt 2: Explore and Commit

4. Attempt 3: Upper Confidence Bound (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,

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

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,

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

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:

Recall the Tool for Building Confidence Interval:

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

$$|\hat{\mu}_k(i) - \mu_k| \leq \sqrt{\frac{\ln(KT/\delta)}{N_t(k)}}$$

Recall the Tool for Building Confidence Interval:

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

$$|\hat{\mu}_k(i) - \mu_k| \leq \sqrt{\frac{\ln(KT/\delta)}{N_t(k)}}$$

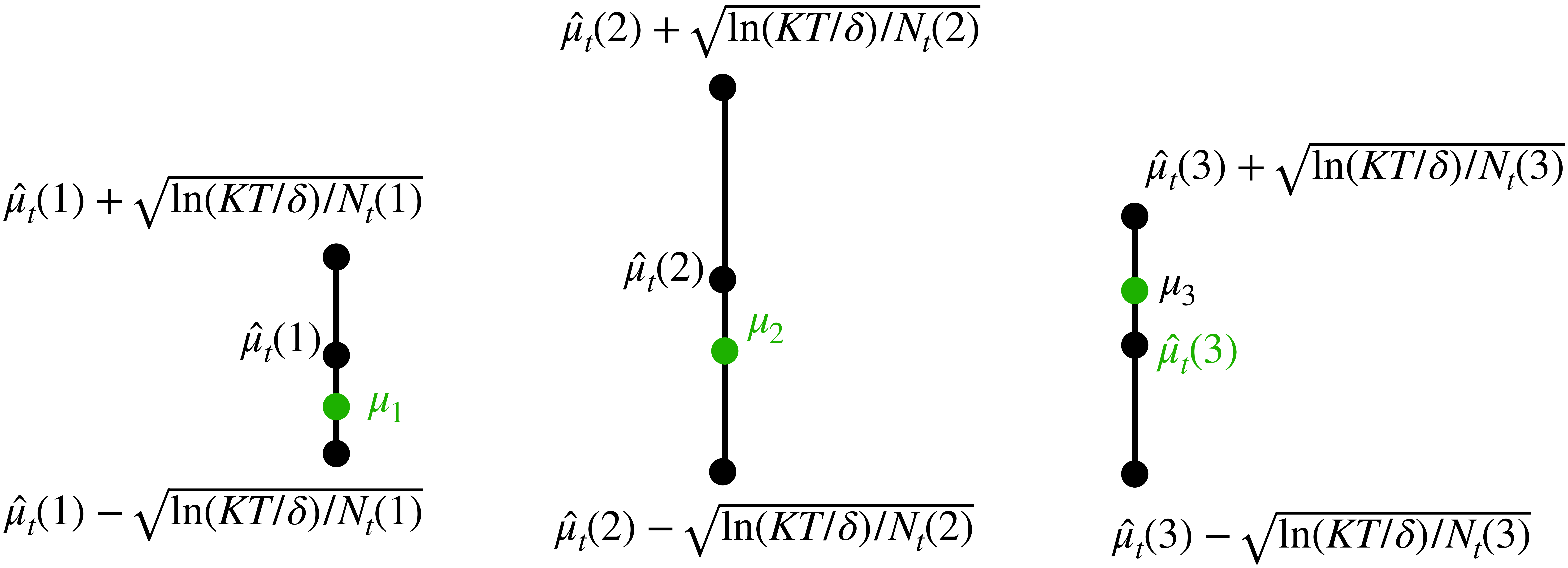
Proving this result actually requires reasoning **Martinalges**, as samples are not i.i.d, i.e., whether or not you pull arm k in this round depends on previous random outcomes (See Ch 6 for more details)

UCB: Optimism in the face of Uncertainty

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

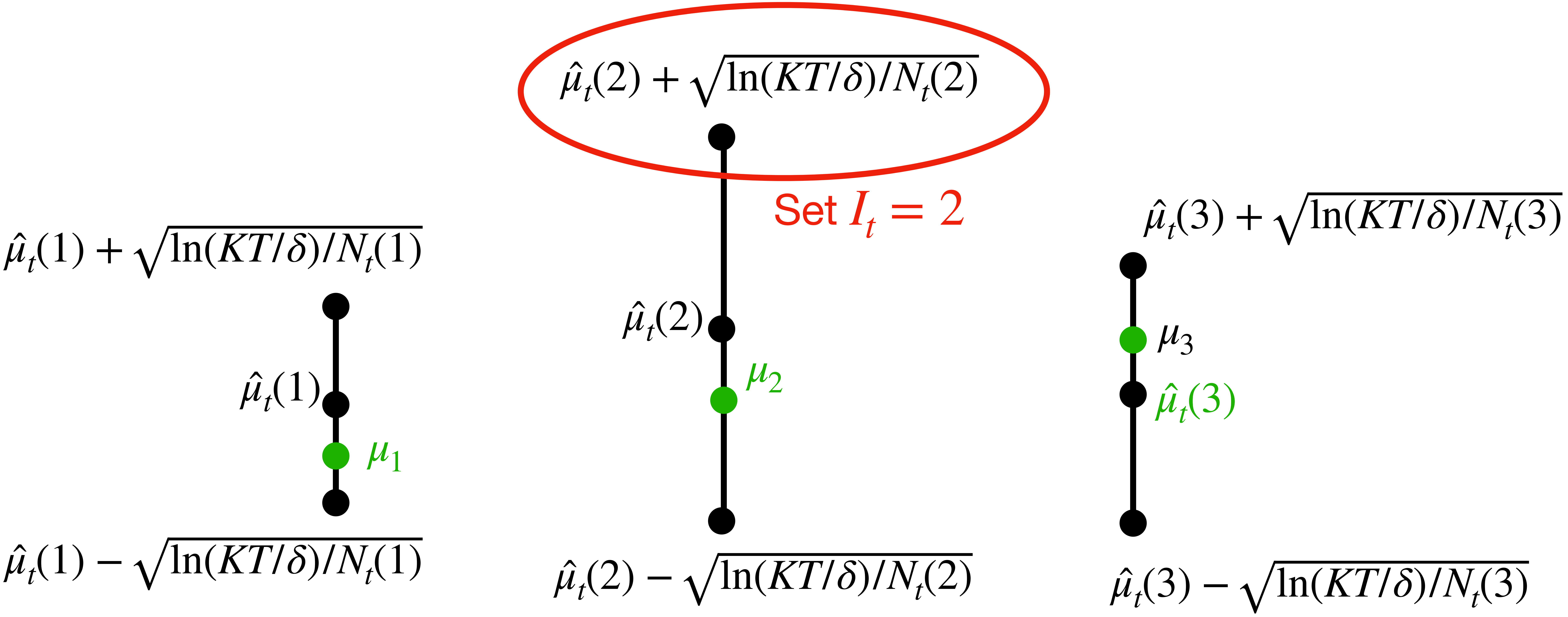
UCB: Optimism in the face of Uncertainty

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 $t = 0 \to T - 1$:

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

Put things together: UCB Algorithm:

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

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

(# 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)

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

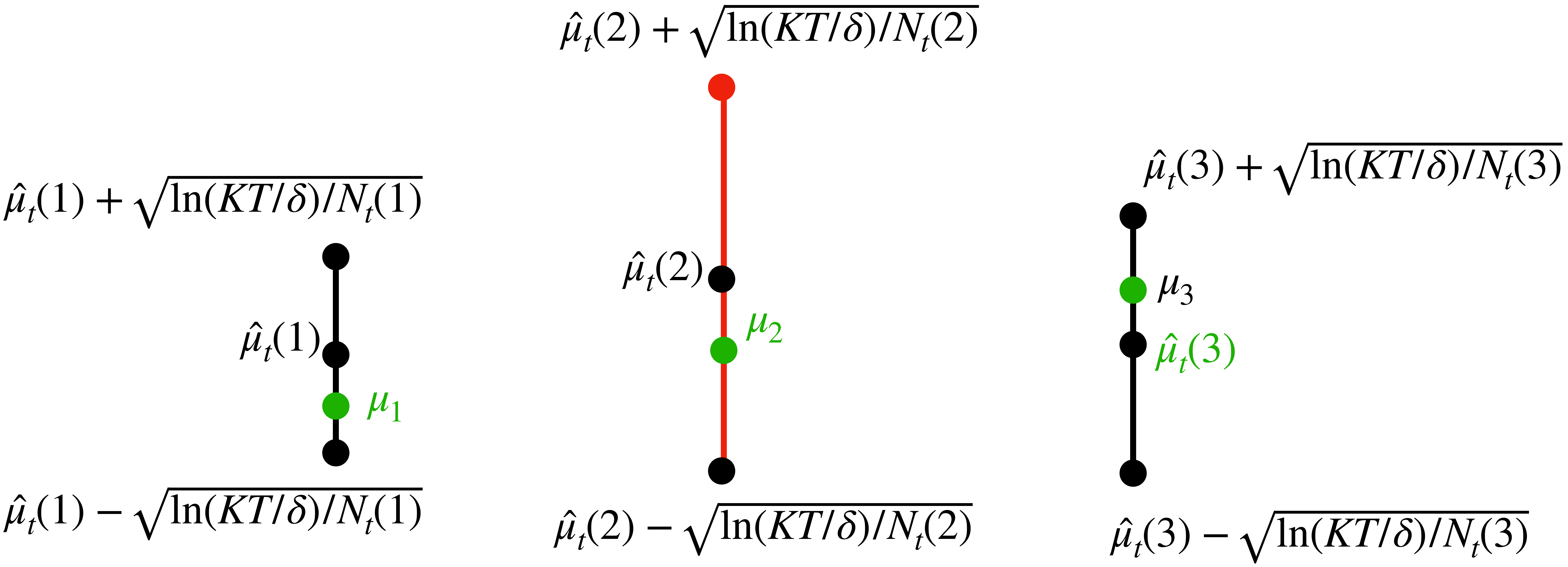
UCB Regret:

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

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

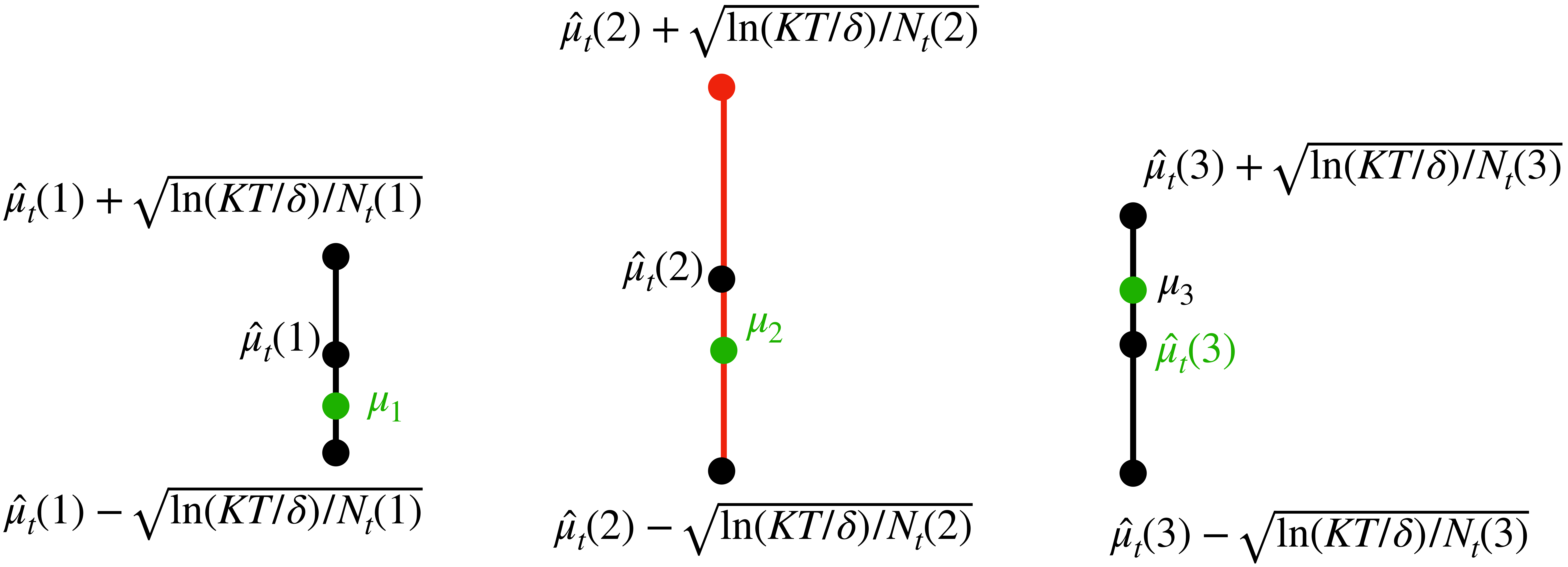
Intuitive Explanation of UCB

Intuitive Explanation of UCB



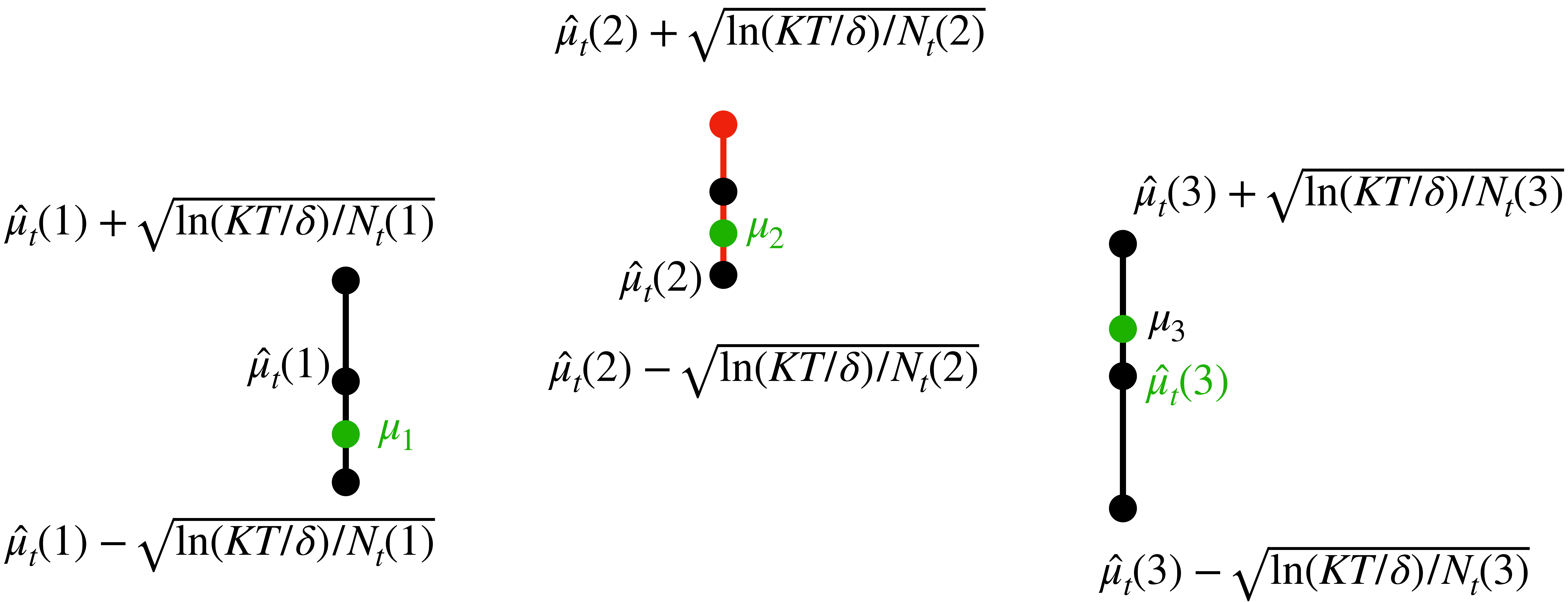
Intuitive Explanation of UCB

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



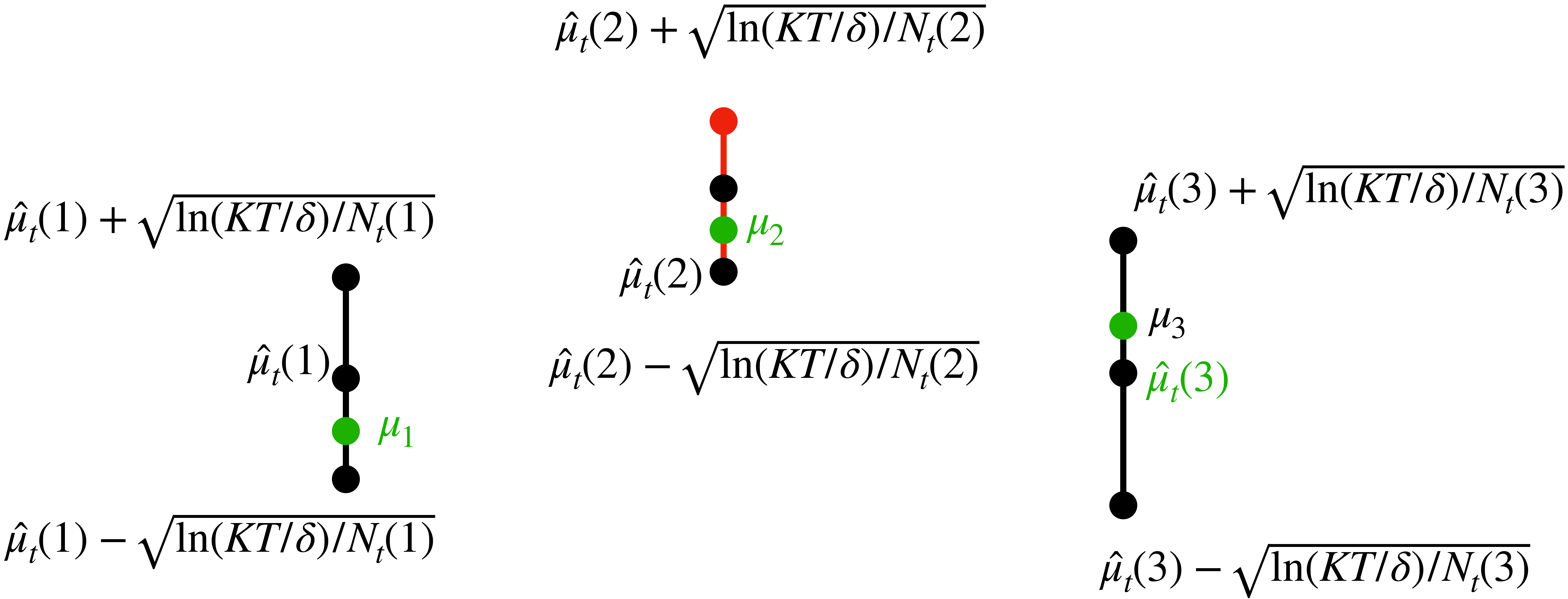
Intuitive Explanation of UCB

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!

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

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

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

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

Let's formalize the intuition

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

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

Q: why?

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

Let's formalize the intuition

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

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

Q: why?

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

$$\leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

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?

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

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

Case 1: $N_t(I_t)$ 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 $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

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: $N_t(I_t)$ 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:

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

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

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

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

Summary

1. Setting of Multi-armed Bandit: MDP with one state, and K actions, H = 1

2. Need to carefully balance exploration and exploitation

3. The Principle of Optimism in the face of Uncertainty