Exploration in RL:
Multi-armed Bandit

Recap: Policy Gradient

The most classic formulation from REINFORCE:

$$\nabla J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho^{\pi_\theta}} \left[R(\tau) \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right]$$

$$\approx R(\tau) \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h), \quad \tau \sim \rho^{\pi_\theta}$$

Recap: Policy Gradient

The most classic formulation from REINFORCE:

$$\nabla J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho^{\pi_\theta}} \left[R(\tau) \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right]$$

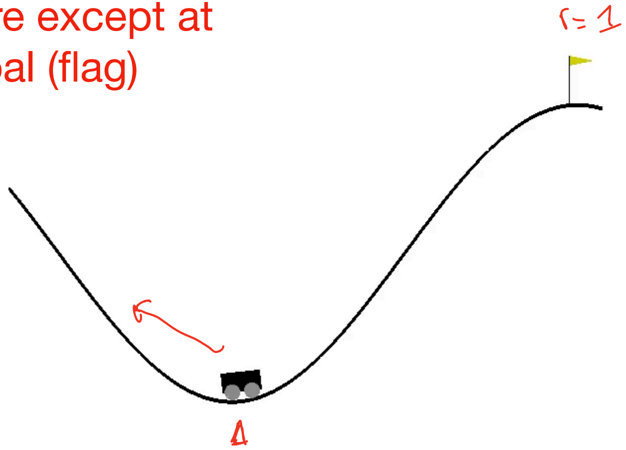
$$\approx R(\tau) \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h), \quad \tau \sim \rho^{\pi_\theta}$$

However, PG lacks the ability to explore;
and it will require much longer time to learn on Acrobot and MountainCar examples in openai Gym

Failure mode of Policy Gradient

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

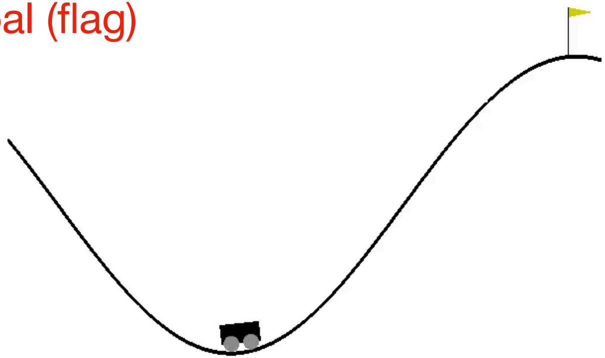
We have reward zero everywhere except at the goal (flag)



Failure mode of Policy Gradient

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

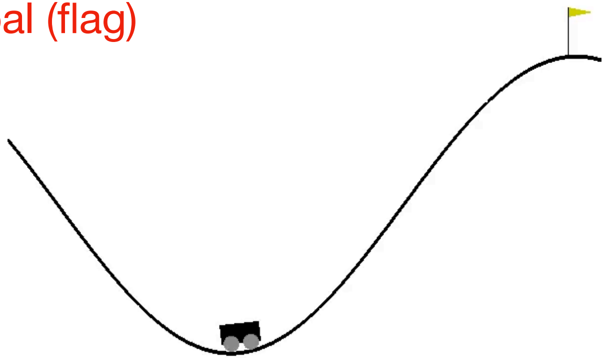
We have reward zero everywhere except at the goal (flag)



Failure mode of Policy Gradient

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

We have reward zero everywhere except at the goal (flag)

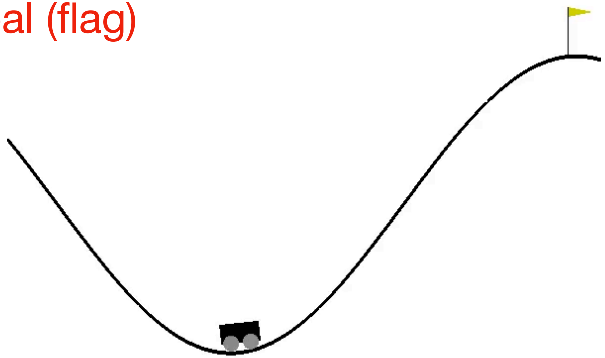


The prob of a random policy hitting the goal is exponentially small $\approx 2^{-H}$

Failure mode of Policy Gradient

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

We have reward zero everywhere except at the goal (flag)



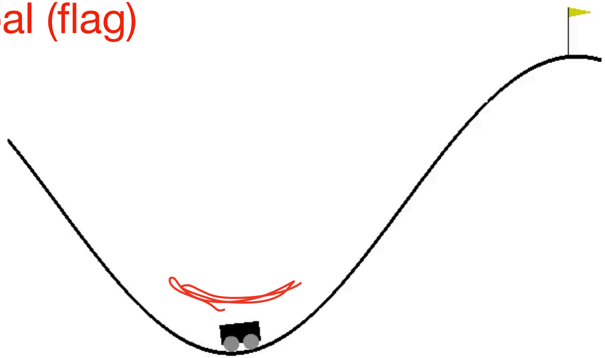
The prob of a random policy hitting the goal is exponentially small $\approx 2^{-H}$

$$\text{PG} := \underbrace{R(\tau)}_{=0} \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \approx 0$$

Failure mode of Policy Gradient

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

We have reward zero everywhere except at the goal (flag)



The prob of a random policy hitting the goal is exponentially small $\approx 2^{-H}$

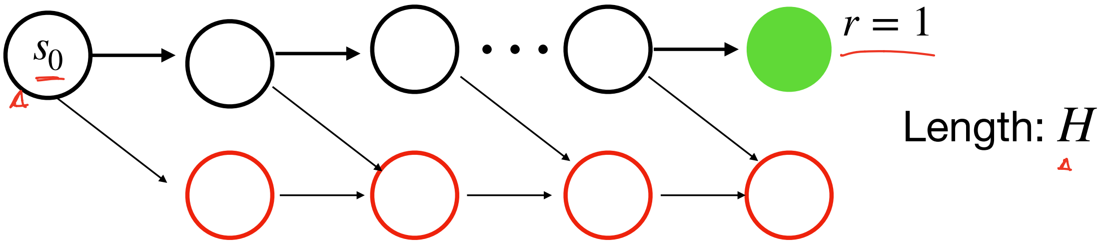
$$\text{PG} := R(\tau) \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \approx 0$$

i.e., a random policy is a perfect locally optimal policy

Failure model of Policy Gradient

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

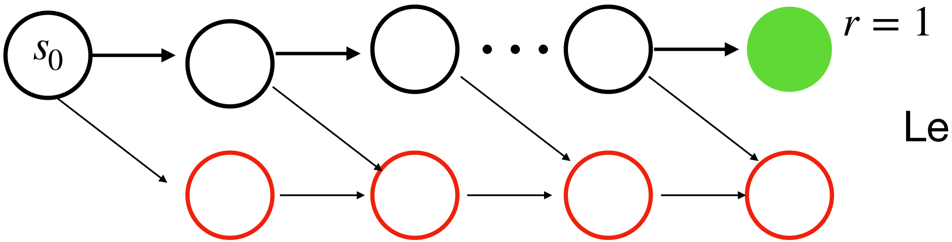
(1) We have reward zero everywhere except at the goal (the right end);
(2) Every black node, one of the two actions will lead the agent to the dead state (red)



Failure model of Policy Gradient

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

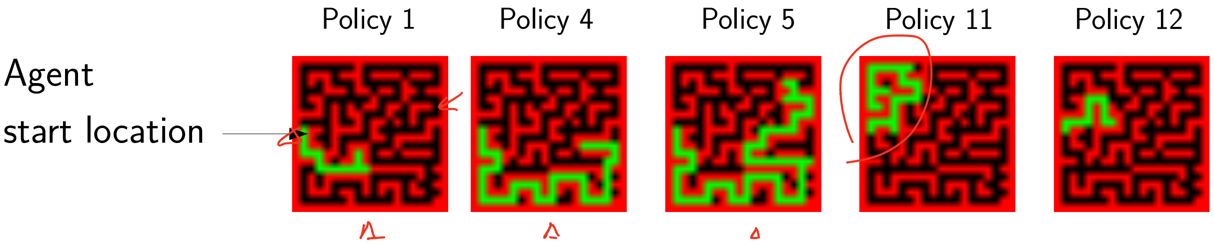
What is the probability of a random policy generating a trajectory that hits the goal?

2^{-H}

Exploration!

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

Exploration in RL is important, but hard...



Example: agent is systematically exploring a maze

Exploration in RL is an active research area, will be treated deeply in CS6789

What we will do here:

Study Exploration in a very simple MDP:

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

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

This is also called Multi-armed Bandits

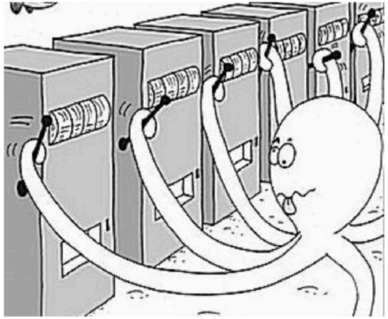
Plan for today:

1. Introduction of MAB

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

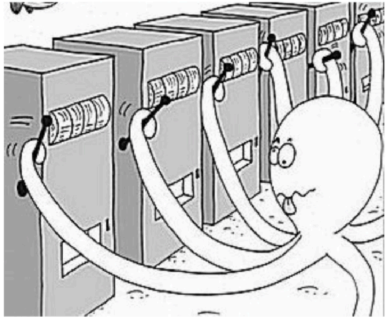
3. Attempt 2: Explore and Commit

Intro to MAB



Setting:

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



Intro to MAB

Setting:

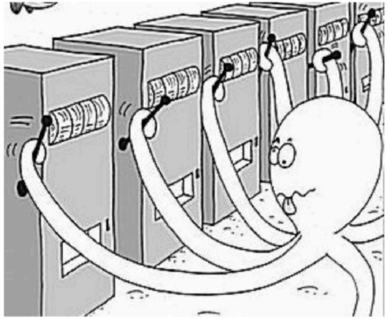
We have K many arms: $a_1, \ldots, a_K$

Each arm has a unknown reward distribution, i.e., $\nu_i \in \Delta([0,1])$,

Δ

w/ mean $\mu_i = \mathbb{E}_{r \sim \nu_i}[r]$

Intro to MAB



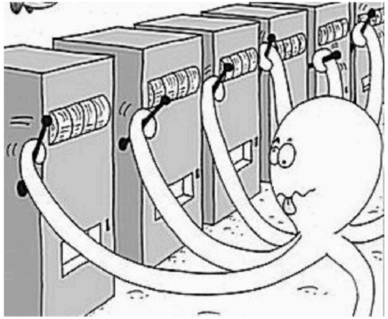
Setting:

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 $t = 0 \rightarrow T - 1$

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

A

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:

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

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

arm we tried at iter $\underline{\underline{t}}$

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

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?

$r \sim \nu_i$

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 $r = 0$ $\mu_1 = 0.6$

Reward dist for a_2: w/ prob 40%, $r = 1$; else $r = 0$ $\mu_2 = 0.4$

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!

40%, $r_1=0$

40%, $r_2=1$

(0.4×0.4)

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**

$0.2(T-2)/T \Rightarrow 0.2$

Plan for today:

1. Introduction of MAB

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

3. Attempt 2: Explore and Commit
- Algorithm
- Analysis

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 $T >> K$)

# of times we will try each arm

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

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 $\{r_i\}_{i=1}^N \overset{i.i.d}{\sim} \nu_k$

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 $\{r_i\}_{i=1}^{N} \sim \nu_k$

μ_k, as $N \rightarrow \infty$

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)

W/K round for exploration

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 $t = \underline{NK} \to T - 1$: (# Exploitation phase)

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: $\hat{\mu}_k = \sum_{i=1}^{N} r_i / N$

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

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

$\mu_1 - - - \mu_K$

T

μ^*

$\frac{T}{K}(\mu_1 - \mu_2)$

$+ \frac{T}{K}(\mu_1 - \mu_3)$

$\vdots$

$+ \frac{T}{K}(\mu_1 - \mu_K)$

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?

Plan for today:

1. Introduction of MAB

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

3. Attempt 2: Explore and Exploit
- Algorithm
- Analysis

Statistical Tools:

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

$\nu \in \Delta[0,1]$ $\quad r_1 \dots r_N \sim \nu$, i.i.d

$\hat{\mu} = r_1 + r_2 + \dots r_N / N$

$|\hat{\mu} - \mu|$, where $\mu = E_{r\sim\nu}[r]$

$\lesssim \sqrt{1/N}$

Statistical Tools:

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

Given a distribution $\overset{\nu}{\cancel{\mu}} \in \Delta([0,1])$, and N i.i.d samples $\{r_i\}_{i=1}^N \sim \overset{\nu}{\cancel{\mu}}$, w/ probability at least $1-\overset{\approx 0.01}{\delta}$, we have:

$$\left|\underbrace{\sum_{i=1}^{N} r_i/N}_{\hat{\mu}} - \underset{\mathbb{E}_{r\sim\nu}[r]}{\mu}\right| \le O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$$

with prob 99%

$$|\hat{\mu}-\mu| \le \sqrt{\frac{5}{N}}$$

$\tilde{O}$

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:

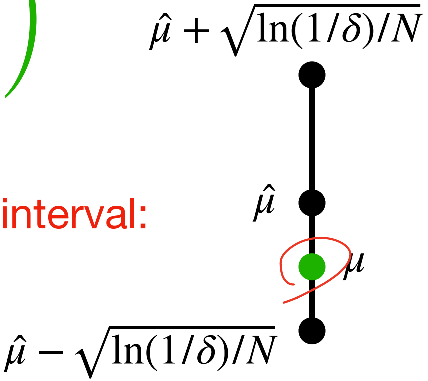
Statistical Tools:

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:

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

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



Statistical Tools:

2. Union Bound (optional):

$$\mathbb{P}\,(A \text{ or } B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

$$P(A_1 \text{ or } A_2 \cdots \text{or } A_k) \leq \sum_{i=1}^{k} P(A_i)$$

area()

$\leq$ area(A) + area(B)

Statistical Tools:

2. Union Bound (optional):

$$\mathbb{P}\left(A \text{ or } B\right) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Combine Hoeffding and Union Bound (optional), we have:

Statistical Tools:

2. Union Bound (optional):

$$\mathbb{P}\left(A \text{ or } B\right) \le \mathbb{P}(A) + \mathbb{P}(B)$$

Combine Hoeffding and Union Bound (optional), we have:

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

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

In summary, we have valid confidence intervals:

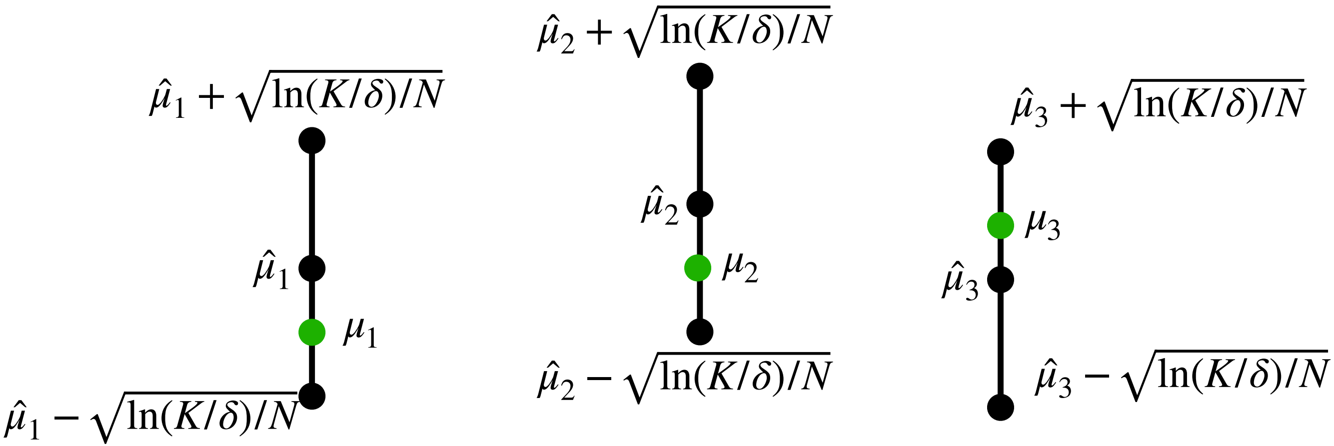
After the Exploration phase, with probability at least 1-δ, **for all arm** $i \in [K]$, we have:

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

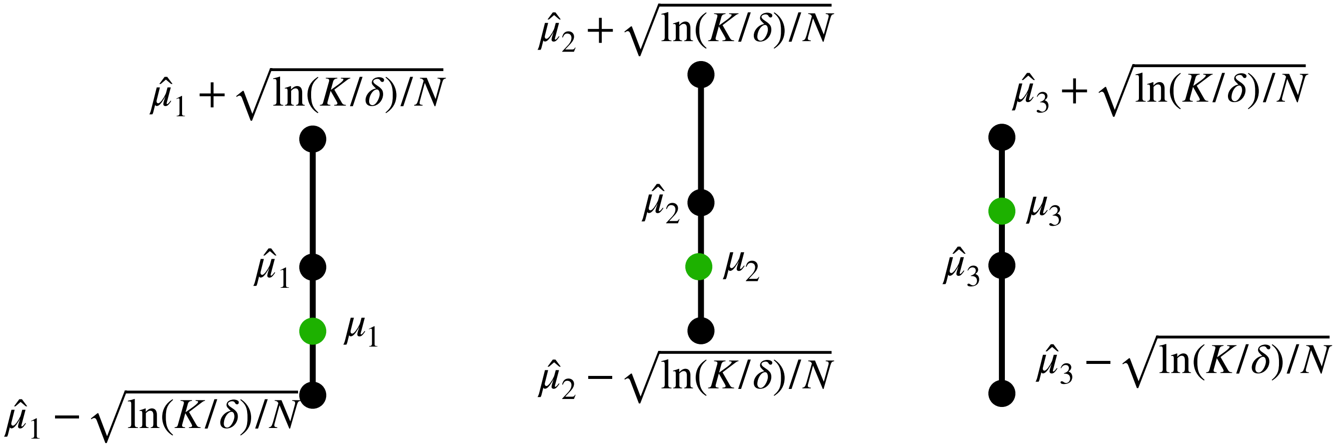
In summary, we have valid confidence intervals:

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

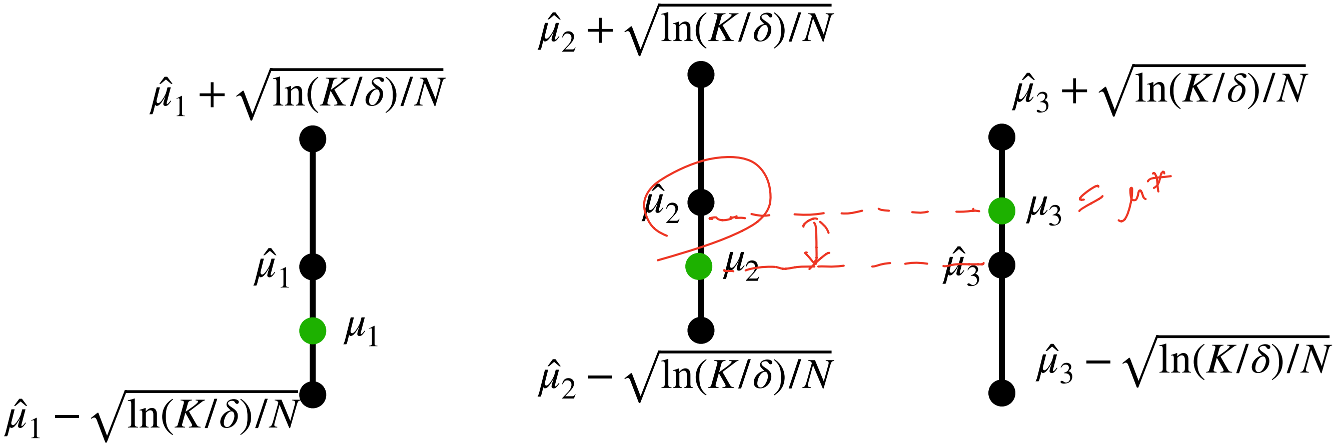
$$\left|\hat{\mu}_i - \mu_i\right| \leq O\left(\sqrt{\frac{\ln(K/\delta)}{N}}\right)$$



In the rest, we will condition on the event that the confidence intervals are valid…



In the rest, we will condition on the event that the confidence intervals are valid…



Recall the Alg in this case will pick $I_t = 2$, for all $t \geq NK$,

(but it will suffer regret $(T - NK)(\mu_3 - \mu_2)$)

Calculate the final regret:

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

our estimations

Ground Truth

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:

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

$\mu_{I^*} - \mu_{\hat{I}} \leq 0$

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:

$\mu_{I^*} \in \left[\hat{\mu}_{I^*} - \sqrt{1/N}, \hat{\mu}_{I^*} + \sqrt{1/N}\right]$

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

$\mu_{\hat{I}} \in \left[\hat{\mu}_{\hat{I}} - \sqrt{1/N}, \hat{\mu}_{\hat{I}} + \sqrt{1/N}\right]$

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

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:

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

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

$$= \hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}} + \underbrace{2\sqrt{\ln(K/\delta)/N}}_{\text{length of confi-Interval}}$$

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:

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

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

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

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

$\forall i \in [K]$,

$\mu_i \in \left[\hat{\mu}_i - \sqrt{\frac{\ln K/\delta}{N}}, \hat{\mu}_i + \sqrt{\frac{\ln K/\delta}{N}}\right]$

$\mu_{I^\star} \leq \hat{\mu}_{I^\star} + \sqrt{\frac{\ln -}{N}}$ (1)

$-\mu_{\hat{I}} \leq -\left[\hat{\mu}_{\hat{I}} - \sqrt{\frac{\ln -}{N}}\right]$ (2)

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

$\hat{\mu}_{\hat{I}} \geq \hat{\mu}_{I^\star}$

What's the regret in the exploitation phase:

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

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

$$= \underbrace{\hat{\mu}_{I^\star} - \hat{\mu}_{\hat{I}}}_{\leq 0} + 2\sqrt{\ln(K/\delta)/N}$$

Q: why?

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

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?

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

$$\approx T\sqrt{\frac{1}{N}}$$

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

Finally, combine two regret together:

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

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

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

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:

of times try each arm

[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:

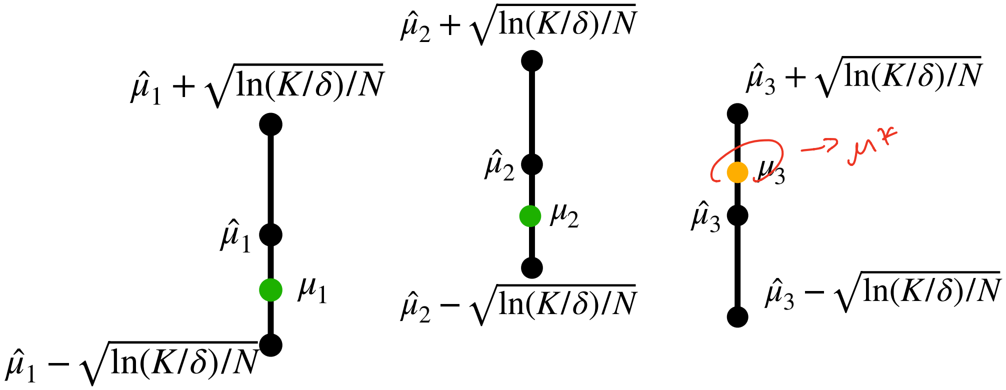
$$\text{Regret}_T \leq O\left(T^{2/3}K^{1/3} \cdot \ln^{1/3}(K/\delta)\right)$$

$\text{Regret}_T / T \approx T^{-1/3} K^{1/3} \to 0$ as $T \to \infty$

(See the MAB reading material for more details)

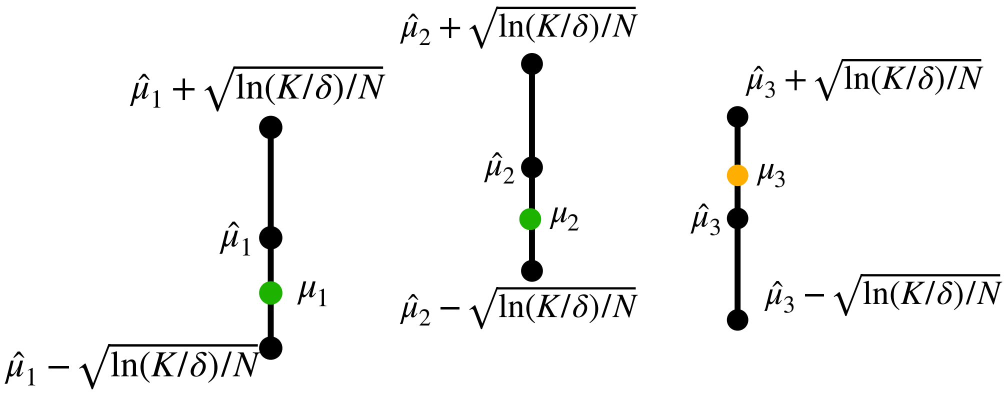
Summary of the analysis:

(1) Using off-shelf statistical tools, we get the confidence intervals for all arms:



Summary of the analysis:

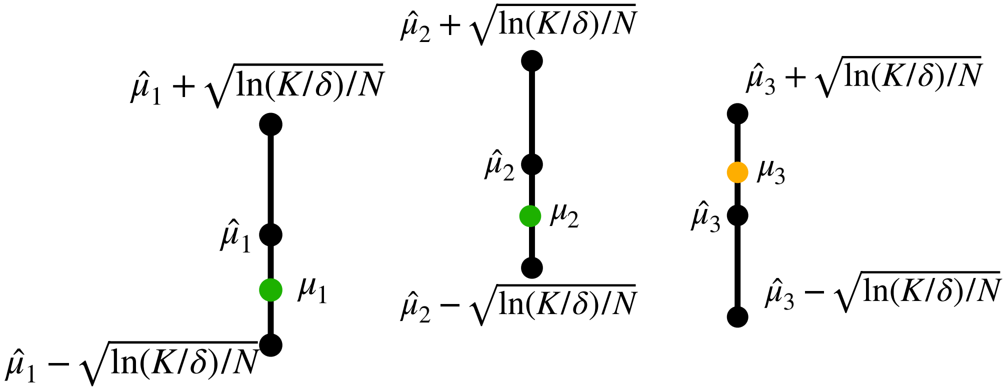
(1) Using off-shelf statistical tools, we get the confidence intervals for all arms:



(2): in the example above, we will commit to arm 2, and pay per-iter regret $\underbrace{(\mu_3 - \mu_2)}_{\leq \text{length-of-Conf-Inter}}$

Summary of the analysis:

(1) Using off-shelf statistical tools, we get the confidence intervals for all arms:



$T^{2/3}$

optimal rate $\sqrt{T}$

(2): in the example above, we will commit to arm 2, and pay per-iter regret $(\mu_3 - \mu_2)$

But from the picture, we see that $(\mu_3 - \mu_2) \leq$ length-of-Confidence-Interval

$\leq \sqrt{1/N}$