Exploration in RL: Multi-armed Bandit (Continue)
Recap: MAB

Interactive learning process:

For $t = 0 \rightarrow T - 1$

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 $I_t$

(# based on historical information)
Recap: MAB

Interactive learning process:

For $t = 0 \rightarrow T - 1$

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:

$$\text{Regret}_T = T \mu^* - \sum_{t=0}^{T-1} \mu_{I_t}$$
Recap: MAB

The Explore and Commit Algorithm:
Recap: MAB

The Explore and Commit Algorithm:

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

The Explore and Commit Algorithm:

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

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

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

1. The upper Confidence Bound Algorithm

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 $i$, 

\[
N_t(i) = \sum_{\tau=0}^{i-1} 1\{I_\tau = i\}
\]
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 $i$,

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

i.e., $N_t(i)$

and its empirical mean $\hat{\mu}_t(i)$ so far;
Statistics that we maintain during learning:

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

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

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

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

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

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

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

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

\[
\hat{\mu}_t(1) + \sqrt{\frac{\ln(KT/\delta)}{N_t(1)}} \\
\hat{\mu}_t(1) - \sqrt{\frac{\ln(KT/\delta)}{N_t(1)}} \\
\hat{\mu}_t(2) + \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \\
\hat{\mu}_t(2) - \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \\
\hat{\mu}_t(3) + \sqrt{\frac{\ln(KT/\delta)}{N_t(3)}} \\
\hat{\mu}_t(3) - \sqrt{\frac{\ln(KT/\delta)}{N_t(3)}}
\]
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**:

\[ \hat{\mu}_t(1) + \sqrt{\ln(KT/\delta)/N_t(1)} \]

\[ \hat{\mu}_t(2) + \sqrt{\ln(KT/\delta)/N_t(2)} \]

\[ \hat{\mu}_t(3) + \sqrt{\ln(KT/\delta)/N_t(3)} \]

\[ \hat{\mu}_t(1) - \sqrt{\ln(KT/\delta)/N_t(1)} \]

\[ \hat{\mu}_t(2) - \sqrt{\ln(KT/\delta)/N_t(2)} \]

\[ \hat{\mu}_t(3) - \sqrt{\ln(KT/\delta)/N_t(3)} \]
UCB: Optimism in the face of Uncertainty

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

\[
\hat{\mu}_t(2) + \sqrt{\ln(KT/\delta)/N_t(2)}
\]

Set \( I_t = 2 \)

\[
\hat{\mu}_t(1) + \sqrt{\ln(KT/\delta)/N_t(1)}
\]

\[
\hat{\mu}_t(1) - \sqrt{\ln(KT/\delta)/N_t(1)}
\]

\[
\hat{\mu}_t(2) - \sqrt{\ln(KT/\delta)/N_t(2)}
\]

\[
\hat{\mu}_t(3) - \sqrt{\ln(KT/\delta)/N_t(3)}
\]

\[
\hat{\mu}_t(3) + \sqrt{\ln(KT/\delta)/N_t(3)}
\]
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 $t = 0 \rightarrow T - 1$:

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

(# 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:

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”: $\sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$
Outline:

1. The upper Confidence Bound Algorithm

2. Analysis of UCB algorithm
Intuitive Explanation of UCB
Intuitive Explanation of UCB

\[ \hat{\mu}_t(1) + \sqrt{\frac{\ln(KT/\delta)}{N_t(1)}} \]

\[ \hat{\mu}_t(1) - \sqrt{\frac{\ln(KT/\delta)}{N_t(1)}} \]

\[ \hat{\mu}_t(2) + \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \]

\[ \hat{\mu}_t(2) - \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \]

\[ \hat{\mu}_t(3) + \sqrt{\frac{\ln(KT/\delta)}{N_t(3)}} \]

\[ \hat{\mu}_t(3) - \sqrt{\frac{\ln(KT/\delta)}{N_t(3)}} \]
Intuitive Explanation of UCB

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

\[
\hat{\mu}_t(1) + \sqrt{\ln(KT/\delta)/N_t(1)}
\]

\[
\hat{\mu}_t(1) - \sqrt{\ln(KT/\delta)/N_t(1)}
\]

\[
\hat{\mu}_t(2) + \sqrt{\ln(KT/\delta)/N_t(2)}
\]

\[
\hat{\mu}_t(2) - \sqrt{\ln(KT/\delta)/N_t(2)}
\]

\[
\hat{\mu}_t(3) + \sqrt{\ln(KT/\delta)/N_t(3)}
\]

\[
\hat{\mu}_t(3) - \sqrt{\ln(KT/\delta)/N_t(3)}
\]
Intuitive Explanation of UCB
Intuitive Explanation of UCB

\[ \hat{\mu}_t(1) + \sqrt{\ln(KT/\delta)/N_t(1)} \]
\[ \hat{\mu}_t(1) - \sqrt{\ln(KT/\delta)/N_t(1)} \]
\[ \hat{\mu}_t(2) + \sqrt{\ln(KT/\delta)/N_t(2)} \]
\[ \hat{\mu}_t(2) - \sqrt{\ln(KT/\delta)/N_t(2)} \]
\[ \hat{\mu}_t(3) + \sqrt{\ln(KT/\delta)/N_t(3)} \]
\[ \hat{\mu}_t(3) - \sqrt{\ln(KT/\delta)/N_t(3)} \]
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!

\[ \hat{\mu}_t(2) + \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \]

\[ \hat{\mu}_t(2) - \sqrt{\frac{\ln(KT/\delta)}{N_t(2)}} \]
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 $I^* = \arg \max_{i \in [K]} \mu_i$; 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 $I^* = \arg \max_{i \in [K]} \mu_i$; recall $I_t = \arg \max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-$t = \mu^* - \mu_{I_t}$
Let’s formalize the intuition

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

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

$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$
Let’s formalize the intuition

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

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

Q: why?
Let’s formalize the intuition

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

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

Q: why?

\[ \leq \hat{\mu}_t(I_t) + \sqrt{\ln(TK/\delta) \over N_t(I_t)} - \mu_{I_t} \]

\[ \leq 2\sqrt{\ln(TK/\delta) \over N_t(I_t)} \]
Let’s formalize the intuition

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

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

Q: why?

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

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

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

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

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 $\mu^*$ & $\mu_{I_t}$ is small)!
Let’s formalize the intuition

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

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

\[
\ln\left( \frac{TK}{\delta} \right) \leq \sqrt{N_t(I_t)}
\]

\[
\leq 2\sqrt{\ln(\frac{TK}{\delta})} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}
\]
Let’s formalize the intuition

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

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

\[ \leq 2^{T-1} \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \]

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

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:

\[ \text{Regret}_T = \tilde{O}\left(\sqrt{KT}\right) \]

(See reading material for more details)
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., exploit)