CS 747, Autumn 2023: Lecture 5

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2023

Multi-armed Bandits

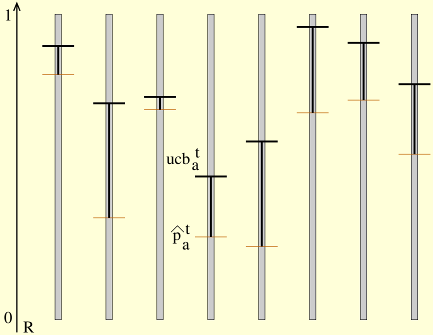
1. Analysis of UCB
2. Other bandit problems

Multi-armed Bandits

1. Analysis of UCB
2. Other bandit problems

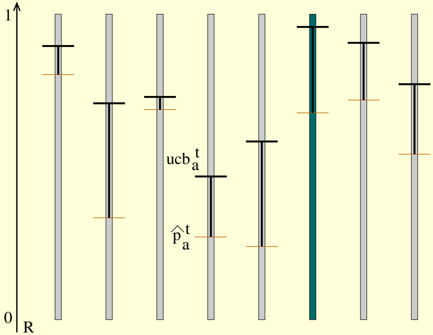
UCB (Auer *et al.*, 2002)

- Pull each arm once.
- For $t \in \{n, n+1, \ldots\}$, for $a \in A$, $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



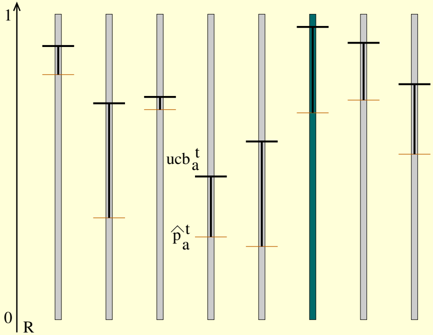
UCB (Auer *et al.*, 2002)

- Pull each arm once.
- For $t \in \{n, n+1, \dots\}$, for $a \in A$, $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



UCB (Auer *et al.*, 2002)

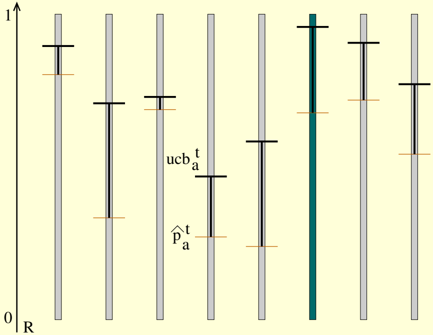
- Pull each arm once.
- For $t \in \{n, n+1, \ldots\}$, for $a \in A$, $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_{a \in A} \text{ucb}_a^t$.



- Recall that $R_T = Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t]$.

UCB (Auer *et al.*, 2002)

- Pull each arm once.
- For $t \in \{n, n+1, \ldots\}$, for $a \in A$, $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



- Recall that $R_T = Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t]$.
- We shall show that UCB achieves $R_T = O\left(\sum_{a: p_a \neq p^\star} \frac{1}{p^\star - p_a} \log(T)\right)$.

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.
- Let Z_a^t be the event that arm a is pulled at time t.

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.

- Let Z_a^t be the event that arm a is pulled at time t.
- Let z_a^t be a random variable that takes value 1 if arm a is pulled at time t, and 0 otherwise.

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.

- Let Z_a^t be the event that arm a is pulled at time t.
- Let z_a^t be a random variable that takes value 1 if arm a is pulled at time t, and 0 otherwise.
  Observe that $\mathbb{E}[z_a^t] = \mathbb{P}\{Z_a^t\}(1) + (1 - \mathbb{P}\{Z_a^t\})(0) = \mathbb{P}\{Z_a^t\}$.

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.

- Let Z_a^t be the event that arm a is pulled at time t.
- Let z_a^t be a random variable that takes value 1 if arm a is pulled at time t, and 0 otherwise.
  Observe that $\mathbb{E}[z_a^t] = \mathbb{P}\{Z_a^t\}(1) + (1 - \mathbb{P}\{Z_a^t\})(0) = \mathbb{P}\{Z_a^t\}$.

- As in the algorithm, u_a^t is a random variable that denotes the number of pulls arm a has received up to time t:

$$u_a^t = \sum_{i=0}^{t-1} z_a^i.$$

Notation

- $\Delta_a \stackrel{\text{def}}{=} p^\star - p_a$ (instance-specific constant); $\star$ an optimal arm.

- Let Z_a^t be the event that arm a is pulled at time t.
- Let z_a^t be a random variable that takes value 1 if arm a is pulled at time t, and 0 otherwise.
  Observe that $\mathbb{E}[z_a^t] = \mathbb{P}\{Z_a^t\}(1) + (1 - \mathbb{P}\{Z_a^t\})(0) = \mathbb{P}\{Z_a^t\}$.

- As in the algorithm, u_a^t is a random variable that denotes the number of pulls arm a has received up to time t:

$$u_a^t = \sum_{i=0}^{t-1} z_a^i.$$

- We define an instance-specific constant $\bar{u}_a^T \stackrel{\text{def}}{=} \left\lceil \frac{8}{(\Delta_a)^2} \ln(T) \right\rceil$ that will serve in our proof as a "sufficient" number of pulls of arm a for horizon T.

Proof Sketch

- To upper-bound R_T, upper-bound the number of pulls of each sub-optimal arm a.

- Give each such arm a $\bar{u}_a^T$ pulls for free.
- Beyond $\bar{u}_a^T$ pulls, arm a's UCB will have width at most $\Delta_a/2$.

- If a continues to be pulled beyond $\bar{u}_a^T$ pulls, either its empirical mean has deviated by more than $\Delta_a/2$ from its true mean, or $\star$'s UCB has fallen below its true mean.
- Both events above have a low probability—in aggregate at most a constant even if summed over an infinite horizon.

Proof Sketch

- To upper-bound R_T, upper-bound the number of pulls of each sub-optimal arm a.

- Give each such arm a $\bar{u}_a^T$ pulls for free.
- Beyond $\bar{u}_a^T$ pulls, arm a's UCB will have width at most $\Delta_a/2$.

- If a continues to be pulled beyond $\bar{u}_a^T$ pulls, either its empirical mean has deviated by more than $\Delta_a/2$ from its true mean, or $\star$'s UCB has fallen below its true mean.
- Both events above have a low probability—in aggregate at most a constant even if summed over an infinite horizon.

- KL-UCB uses the KL inequality, and slightly more sophisticated analysis.

Step 1: Show that $R_T = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

Step 1: Show that $R_T = \sum_{a:p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$R_T = Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t]$$

Step 1: Show that $R_T = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$R_T = Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\}\mathbb{E}[r^t | Z_a^t]$$

Step 1: Show that $R_T = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$R_T = Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\}\mathbb{E}[r^t | Z_a^t]$$

$$= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a$$

Step 1: Show that $R_T = \sum_{a:p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$
\begin{aligned}
R_T &= Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\} \mathbb{E}[r^t | Z_a^t] \\
&= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T] \right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a
\end{aligned}
$$

Step 1: Show that $R_T = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$
\begin{aligned}
R_T &= Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\} \mathbb{E}[r^t | Z_a^t] \\
&= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T] \right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a \\
&= \sum_{a \in A} \mathbb{E}[u_a^T](p^\star - p_a)
\end{aligned}
$$

Step 1: Show that $R_T = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$
\begin{aligned}
R_T &= Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\}\mathbb{E}[r^t | Z_a^t] \\
&= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T] \right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a \\
&= \sum_{a \in A} \mathbb{E}[u_a^T](p^\star - p_a) = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a.
\end{aligned}
$$

Step 1: Show that $R_T = \sum_{a:p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a$.

$$
\begin{aligned}
R_T &= Tp^\star - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\}\mathbb{E}[r^t | Z_a^t] \\
&= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T]\right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a \\
&= \sum_{a \in A} \mathbb{E}[u_a^T](p^\star - p_a) = \sum_{a:p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a.
\end{aligned}
$$

To show the regret bound, we shall show for each sub-optimal arm a that

$$
\mathbb{E}[u_a^T] = O\left(\frac{1}{(\Delta_a)^2} \log(T)\right).
$$

Step 2: Two Regimes for Sub-optimal Pulls

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

$$\mathbb{E}[u_a^T] = \sum_{t=0}^{T-1} \mathbb{E}[z_a^t]$$

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

$$\mathbb{E}[u_a^T] = \sum_{t=0}^{T-1} \mathbb{E}[z_a^t] = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t\}$$

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

$$\begin{aligned}\mathbb{E}[u_a^T] &= \sum_{t=0}^{T-1}\mathbb{E}[z_a^t] = \sum_{t=0}^{T-1}\mathbb{P}\{Z_a^t\}\\ &= \sum_{t=0}^{T-1}\mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\} + \sum_{t=0}^{T-1}\mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}\end{aligned}$$

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

$$\begin{aligned}
\mathbb{E}[u_a^T] &= \sum_{t=0}^{T-1} \mathbb{E}[z_a^t] = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t\} \\
&= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\} + \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&= A + B.
\end{aligned}$$

Step 2: Two Regimes for Sub-optimal Pulls

To prove $\mathbb{E}[u_a^T] = O\left(\frac{1}{\Delta_a^2}\log(T)\right)$, we show $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ for constant C.

$$\begin{aligned}
\mathbb{E}[u_a^T] &= \sum_{t=0}^{T-1} \mathbb{E}[z_a^t] = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t\} \\
&= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\} + \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&= A + B.
\end{aligned}$$

We show A is upper-bounded by $\bar{u}_a^T$ and B is upper-bounded by a constant.

Step 3: Bounding A

Step 3: Bounding A

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

Step 3: Bounding A

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

$$= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\}$$

Step 3: Bounding A

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

$$= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} = \sum_{m=0}^{\bar{u}_a^T - 1} \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\}$$

Step 3: Bounding A

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

$$= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} = \sum_{m=0}^{\bar{u}_a^T - 1} \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\}$$

$$= \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^0, (u_a^0 = m) \text{ or } Z_a^1, (u_a^1 = m) \text{ or } \ldots \text{ or } \; Z_a^{T-1}, (u_a^{T-1} = m)\}$$

Step 3: Bounding A

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

$$= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} = \sum_{m=0}^{\bar{u}_a^T - 1} \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\}$$

$$= \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^0, (u_a^0 = m) \text{ or } Z_a^1, (u_a^1 = m) \text{ or } \ldots \text{ or } \ Z_a^{T-1}, (u_a^{T-1} = m)\}$$

$$\leq \sum_{m=0}^{\bar{u}_a^T - 1} 1$$

Step 3: Bounding A

$$
\begin{aligned}
A &= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}\\
&= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} = \sum_{m=0}^{\bar{u}_a^T-1} \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\}\\
&= \sum_{m=0}^{\bar{u}_a^T-1} \mathbb{P}\{Z_a^0, (u_a^0 = m) \text{ or } Z_a^1, (u_a^1 = m) \text{ or } \ldots \text{ or } \; Z_a^{T-1}, (u_a^{T-1} = m)\}\\
&\leq \sum_{m=0}^{\bar{u}_a^T-1} 1 = \bar{u}_a^T.
\end{aligned}
$$

Step 3: Bounding A

$$
\begin{aligned}
A &= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\} \\
&= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} = \sum_{m=0}^{\bar{u}_a^T - 1} \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t = m)\} \\
&= \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^0, (u_a^0 = m) \text{ or } Z_a^1, (u_a^1 = m) \text{ or } \ldots \text{ or } \ Z_a^{T-1}, (u_a^{T-1} = m)\} \\
&\leq \sum_{m=0}^{\bar{u}_a^T - 1} 1 = \bar{u}_a^T.
\end{aligned}
$$

We have used the fact that for $0 \leq i < j \leq t-1$, $(Z_a^i, (u_a^i = m))$ and $(Z_a^j, (u_a^j = m))$ are mutually exclusive.

Step 4.1: Bounding B

Step 4.1: Bounding B

$$B = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

Step 4.1: Bounding B

$$B = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$= \sum_{t=n}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

Step 4.1: Bounding B

$$B = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$= \sum_{t=n}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$\leq \sum_{t=n}^{T-1} \mathbb{P}\left\{\left(\hat{p}_a^t + \sqrt{\frac{2}{u_a^t}\ln(t)} \geq \hat{p}_\star^t + \sqrt{\frac{2}{u_\star^t}\ln(t)}\right) \text{ and } (u_a^t \geq \bar{u}_a^T)\right\}$$

Step 4.1: Bounding B

$$B = \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$= \sum_{t=n}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$\leq \sum_{t=n}^{T-1} \mathbb{P}\left\{\left(\hat{p}_a^t + \sqrt{\frac{2}{u_a^t}\ln(t)} \geq \hat{p}_\star^t + \sqrt{\frac{2}{u_\star^t}\ln(t)}\right) \text{ and } (u_a^t \geq \bar{u}_a^T)\right\}$$

$$\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\} \text{ where}$$

$\hat{p}_a(x)$ is the empirical mean of the first x pulls of arm a, and
$\hat{p}_\star(y)$ is the empirical mean of the first y pulls of arm $\star$.

Step 4.2: Bounding B

- Fix $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \dots, t\}$ and $y \in \{1, 2, \dots, t\}$.

Step 4.2: Bounding B

- Fix $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \ldots, t\}$ and $y \in \{1, 2, \ldots, t\}$.

1. We have:

$$\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}$$

$$\implies \left(\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq p_\star\right) \text{ or } \left(\hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)} < p_\star\right).$$

Step 4.2: Bounding B

- Fix $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \ldots, t\}$ and $y \in \{1, 2, \ldots, t\}$.

1. We have:

$$\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}$$

$$\implies \left(\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq p_\star\right) \text{ or } \left(\hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)} < p_\star\right).$$

Fact: If $\alpha > \beta$, then $\alpha \geq \gamma$ or $\beta < \gamma$. Holds for arbitrary α, β, γ!

Step 4.2: Bounding B

- Fix $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \ldots, t\}$ and $y \in \{1, 2, \ldots, t\}$.

1. We have:

$$\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}$$

$$\implies \left(\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq p_\star\right) \text{ or } \left(\hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)} < p_\star\right).$$

   Fact: If $\alpha > \beta$, then $\alpha \geq \gamma$ or $\beta < \gamma$. Holds for arbitrary α, β, γ!

2. Since $x \geq \bar{u}_a^T$, we have $\sqrt{\frac{2}{x}\ln(t)} \leq \sqrt{\frac{2}{\bar{u}_a^T}\ln(t)} \leq \frac{\Delta_a}{2}$, and so

$$\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq p_\star \implies \hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}.$$

Step 4.3: Bounding B

Continuing from Step 4.1, using the two results from Step 4.2, and invoking Hoeffding's Inequality:

$$B \leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

Step 4.3: Bounding B

Continuing from Step 4.1, using the two results from Step 4.2, and invoking Hoeffding's Inequality:

$$B \leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

Step 4.3: Bounding B

Continuing from Step 4.1, using the two results from Step 4.2, and invoking Hoeffding's Inequality:

$$\begin{aligned}
B &\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\} \\
&\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right) \\
&\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right)
\end{aligned}$$

Step 4.3: Bounding B

Continuing from Step 4.1, using the two results from Step 4.2, and invoking Hoeffding's Inequality:

$$
\begin{aligned}
B &\le \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \ge \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\} \\
&\le \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \ge p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right) \\
&\le \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right) \\
&\le \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-4\ln(t)} + e^{-4\ln(t)}\right) \le \sum_{t=n}^{T-1} t^2\left(\frac{2}{t^4}\right) \le \sum_{t=1}^{\infty} \frac{2}{t^2} = \frac{\pi^2}{3}.
\end{aligned}
$$

Step 4.3: Bounding B

Continuing from Step 4.1, using the two results from Step 4.2, and invoking Hoeffding's Inequality:

$$B \leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

$$\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right)$$

$$\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-4\ln(t)} + e^{-4\ln(t)}\right) \leq \sum_{t=n}^{T-1} t^2 \left(\frac{2}{t^4}\right) \leq \sum_{t=1}^{\infty} \frac{2}{t^2} = \frac{\pi^2}{3}.$$

We are done!

Multi-armed Bandits

1. Analysis of UCB

2. Other bandit problems

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

- Incorporating risk/variance in the objective.
  - Arm 1 gives rewards 0 and 100, each w.p. $1/2$.
  - Arm 2 gives rewards 48 and 50, each w.p. $1/2$.
  - Which arm would you prefer?

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

- Incorporating risk/variance in the objective.
  - Arm 1 gives rewards 0 and 100, each w.p. $1/2$.
  - Arm 2 gives rewards 48 and 50, each w.p. $1/2$.
  - Which arm would you prefer?

- What if the arms' (true) means vary over time?
  - Nonstationary setting, seen for example, in on-line ads.
  - Approach depends on nature of drift/change in rewards.
  - In practice, one might only trust most recent data from arms.
  - In practice, the set of arms can itself change over time!

Other Bandit Problems

- Pure exploration.
  - Separate "testing" and "live" phases.
  - In testing phase, rewards don't matter.
  - PAC formulation: W.p. at least $1 - \delta$, must return an ϵ-optimal arm, while incurring a small number of pulls.
  - Simple regret formulation: Given a budget of T pulls, must output an arm a such that p_a is large, or equivalently, simple regret $= p^\star - p_a$ is small).

Other Bandit Problems

- Pure exploration.
  - Separate “testing” and “live” phases.
  - In testing phase, rewards don’t matter.
  - PAC formulation: W.p. at least $1 - \delta$, must return an ϵ-optimal arm, while incurring a small number of pulls.
  - Simple regret formulation: Given a budget of T pulls, must output an arm a such that p_a is large, or equivalently, simple regret $= p^\star - p_a$ is small).

- Limited number of feedback stages.
  - Suppose you are given budget T, but your algorithm can look at history only $s < T$ times?
  - UCB, Thompson Sampling, etc. are fully sequential ($s = T$).
  - How to manage with fewer “stages” s?

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for "good", rather than "optimal" arms.

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for “good”, rather than “optimal” arms.
- What if we are interacting with many bandits simultaneously?
  - Contextual bandits: If the bandits themselves can be described using features (a “context”), data from one can be used to generate estimates about others.

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for “good”, rather than “optimal” arms.

- What if we are interacting with many bandits simultaneously?
  - Contextual bandits: If the bandits themselves can be described using features (a “context”), data from one can be used to generate estimates about others.

- What if the rewards do not come from a fixed random process?
  - Adversarial bandits make no assumption on the rewards.
  - Possible to show sub-linear regret when compared against playing a single arm for the entire run.
  - Necessary to use a randomised algorithm.

Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- ϵ-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm
- Concentration bounds
- Understanding Thompson Sampling
- Other bandit problems

Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- ϵ-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm
- Concentration bounds
- Understanding Thompson Sampling
- Other bandit problems

- **Next class:** Markov Decision Problems