CS 747, Autumn 2022: Lecture 4

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB

Multi-armed Bandits

1. Concentration bounds

2. Analysis of UCB

Hoeffding's Inequality (Hoeffding, 1963)

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;

Hoeffding's Inequality (Hoeffding, 1963)

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and

Hoeffding's Inequality (Hoeffding, 1963)

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and
- Let $\bar{x}$ be the mean of these samples (an *empirical* mean):

$$\bar{x} = \frac{1}{u} \sum_{i=1}^{u} x_i.$$

Hoeffding's Inequality (Hoeffding, 1963)

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and
- Let $\bar{x}$ be the mean of these samples (an *empirical* mean):

$$\bar{x} = \frac{1}{u} \sum_{i=1}^{u} x_i.$$

- Then, for or any fixed $\epsilon > 0$, we have

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u\epsilon^2}, \text{ and}$$
$$\mathbb{P}\{\bar{x} \leq \mu - \epsilon\} \leq e^{-2u\epsilon^2}.$$

Hoeffding's Inequality (Hoeffding, 1963)

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and
- Let $\bar{x}$ be the mean of these samples (an *empirical* mean):

$$\bar{x} = \frac{1}{u} \sum_{i=1}^{u} x_i.$$

- Then, for or any fixed $\epsilon > 0$, we have

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u\epsilon^2}, \text{ and}$$
$$\mathbb{P}\{\bar{x} \leq \mu - \epsilon\} \leq e^{-2u\epsilon^2}.$$

- Note the bounds are trivial for large ϵ, since $\bar{x} \in [0, 1]$.

Applications

- For given mistake probability δ and tolerance ϵ, how many samples u_0 of X do we need to guarantee that with probability at least $1 - \delta$, the empirical mean $\bar{x}$ will not exceed the true mean μ by ϵ or more?

Applications

- For given mistake probability δ and tolerance ϵ, how many samples u_0 of X do we need to guarantee that with probability at least $1 - \delta$, the empirical mean $\bar{x}$ will not exceed the true mean μ by ϵ or more?

  $u_0 = \lceil \frac{1}{2\epsilon^2} \ln(\frac{1}{\delta}) \rceil$ pulls are sufficient, since Hoeffding's Inequality gives

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u_0\epsilon^2} \leq \delta.$$

Applications

- For given mistake probability δ and tolerance ϵ, how many samples u_0 of X do we need to guarantee that with probability at least $1 - \delta$, the empirical mean $\bar{x}$ will not exceed the true mean μ by ϵ or more?
  $u_0 = \lceil \frac{1}{2\epsilon^2} \ln(\frac{1}{\delta}) \rceil$ pulls are sufficient, since Hoeffding's Inequality gives

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u_0\epsilon^2} \leq \delta.$$

- We have u samples of X. How do we fill up this blank?:
  With probability at least $1 - \delta$, the empirical mean $\bar{x}$ exceeds the true mean μ by at most $\epsilon_0 =$ __________.

Applications

- For given mistake probability δ and tolerance ϵ, how many samples u_0 of X do we need to guarantee that with probability at least $1 - \delta$, the empirical mean $\bar{x}$ will not exceed the true mean μ by ϵ or more?
  $u_0 = \lceil \frac{1}{2\epsilon^2} \ln(\frac{1}{\delta}) \rceil$ pulls are sufficient, since Hoeffding's Inequality gives

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u_0\epsilon^2} \leq \delta.$$

- We have u samples of X. How do we fill up this blank?:
  With probability at least $1 - \delta$, the empirical mean $\bar{x}$ exceeds the true mean μ by at most $\epsilon_0 =$ __________.
  We can write $\epsilon_0 = \sqrt{\frac{1}{2u} \ln(\frac{1}{\delta})}$; by Hoeffding's Inequality:

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon_0\} \leq e^{-2u(\epsilon_0)^2} \leq \delta.$$

Arbitrary Bounded Range

- Suppose X is a random variable bounded in $[a, b]$. Can we still apply Hoeffding's Inequality?

Arbitrary Bounded Range

- Suppose X is a random variable bounded in $[a, b]$. Can we still apply Hoeffding's Inequality?

  Yes. Assume u; $x_1, x_2, \ldots, x_u$; ϵ as defined earlier.

Arbitrary Bounded Range

- Suppose X is a random variable bounded in $[a, b]$. Can we still apply Hoeffding's Inequality?

  Yes. Assume $u; x_1, x_2, \ldots, x_u; \epsilon$ as defined earlier.

  Consider $Y = \frac{X-a}{b-a}$; for $1 \leq i \leq u$, $y_i = \frac{x_i - a}{b-a}$; $\bar{y} = \frac{1}{u}\sum_{i=1}^{u} y_i$.

Arbitrary Bounded Range

- Suppose X is a random variable bounded in $[a, b]$. Can we still apply Hoeffding's Inequality?

  Yes. Assume u; $x_1, x_2, \ldots, x_u$; ϵ as defined earlier.

  Consider $Y = \frac{X-a}{b-a}$; for $1 \leq i \leq u$, $y_i = \frac{x_i - a}{b-a}$; $\bar{y} = \frac{1}{u}\sum_{i=1}^{u} y_i$.

  Since Y is bounded in $[0, 1]$, we get

$$\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} = \mathbb{P}\left\{\bar{y} \geq \frac{\mu - a}{b - a} + \frac{\epsilon}{b - a}\right\} \leq e^{-\frac{2u\epsilon^2}{(b-a)^2}}, \text{ and}$$

$$\mathbb{P}\{\bar{x} \leq \mu - \epsilon\} = \mathbb{P}\left\{\bar{y} \leq \frac{\mu - a}{b - a} - \frac{\epsilon}{b - a}\right\} \leq e^{-\frac{2u\epsilon^2}{(b-a)^2}}.$$

A “KL” Inequality

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and
- Let $\bar{x}$ be the mean of these samples (an *empirical* mean):

$$\bar{x} = \frac{1}{u} \sum_{i=1}^{u} x_i.$$

A "KL" Inequality

- Let X be a random variable bounded in $[0, 1]$, with $\mathbb{E}[X] = \mu$;
- Let $u \geq 1$;
- Let $x_1, x_2, \ldots, x_u$ be i.i.d. samples of X; and
- Let $\bar{x}$ be the mean of these samples (an *empirical* mean):

$$\bar{x} = \frac{1}{u} \sum_{i=1}^{u} x_i.$$

- Then, for or any fixed $\epsilon \in [0, 1 - \mu]$, we have

$$\boxed{\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-uKL(\mu+\epsilon,\mu)},}$$

and for or any fixed $\epsilon \in [0, \mu]$, we have

$$\boxed{\mathbb{P}\{\bar{x} \leq \mu - \epsilon\} \leq e^{-uKL(\mu-\epsilon,\mu)},}$$

where for $p, q \in [0, 1]$, $KL(p, q) \stackrel{\text{def}}{=} p \ln(\frac{p}{q}) + (1 - p) \ln(\frac{1-p}{1-q})$.

Some Observations

- The KL inequality gives a tighter upper bound:
  For $p, q \in [0, 1]$,

$$KL(p, q) \geq 2(p - q)^2 \implies e^{-uKL(p,q)} \leq e^{-2u(p-q)^2}.$$

- Both bounds are instances of "Chernoff bounds", of which there are many more forms.

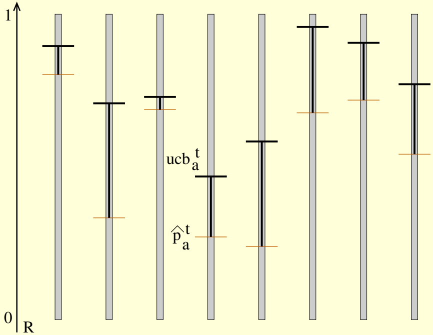
- Similar bounds can also be given when X has infinite support (such as a Gaussian), but might need additional assumptions.

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB

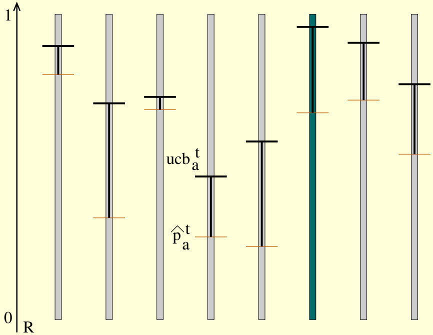
UCB (Auer *et al.*, 2002)

- 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 $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



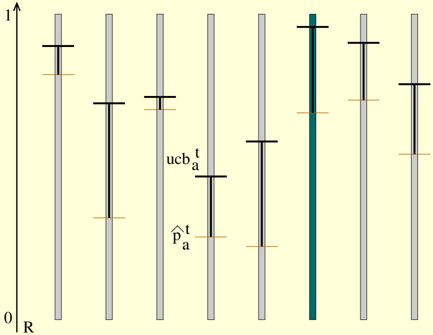
UCB (Auer *et al.*, 2002)

- 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 $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



UCB (Auer *et al.*, 2002)

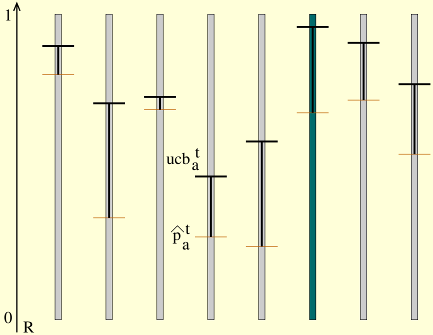
- Pull each arm once.
- For $t \in \{n, n+1, \ldots\}$, for $a \in A$, $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_{a \in A}\, \text{ucb}_a^t$.



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

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

$$= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T] \right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a$$

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

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

$$= Tp^\star - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a = \left(\sum_{a \in A} \mathbb{E}[u_a^T] \right) p^\star - \sum_{a \in A} \mathbb{E}[u_a^T] p_a$$

$$= \sum_{a \in A} \mathbb{E}[u_a^T](p^\star - p_a) = \sum_{a: p_a \neq p^\star} \mathbb{E}[u_a^T]\Delta_a.$$

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.

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

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.

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

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

$$= A + B.$$

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

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

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

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

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

Step 3: Bounding A

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

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

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

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

Step 3: Bounding A

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

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

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

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

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

$$
\begin{aligned}
B &= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&= \sum_{t=n}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&\leq \sum_{t=n}^{T-1} \mathbb{P}\left\{\left(\hat{p}_a^t + \sqrt{\frac{2}{u_a^t}\ln(t)} \geq \hat{p}_\star^t + \sqrt{\frac{2}{u_\star^t}\ln(t)}\right) \text{ and } (u_a^t \geq \bar{u}_a^T)\right\}
\end{aligned}
$$

Step 4.1: Bounding B

$$
\begin{aligned}
B &= \sum_{t=0}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&= \sum_{t=n}^{T-1} \mathbb{P}\{Z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\} \\
&\leq \sum_{t=n}^{T-1} \mathbb{P}\left\{\left(\hat{p}_a^t + \sqrt{\frac{2}{u_a^t}\ln(t)} \geq \hat{p}_\star^t + \sqrt{\frac{2}{u_\star^t}\ln(t)}\right) \text{ and } (u_a^t \geq \bar{u}_a^T)\right\} \\
&\leq \sum_{t=n}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\} \text{ where}
\end{aligned}
$$

$\hat{p}_a(x)$ 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 $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \ldots, t\}$ and $y \in \{1, 2, \ldots, t\}$.

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

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:

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

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

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

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

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!