CS 747, Autumn 2020: Week 3, Lecture 1

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2020

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB
3. Understanding Thompson Sampling
4. Other bandit problems

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB
3. Understanding Thompson Sampling
4. Other bandit problems

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

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

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 $x_1, x_2, \dots, x_u$ 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.

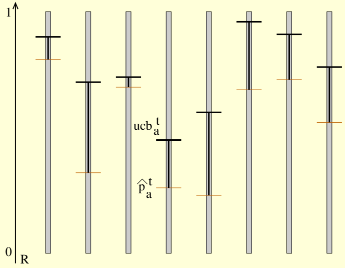
- Similar bounds can also be given when X has infinite support (such as a Gaussian), but might need additional assumptions.

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB
3. Understanding Thompson Sampling
4. Other bandit problems

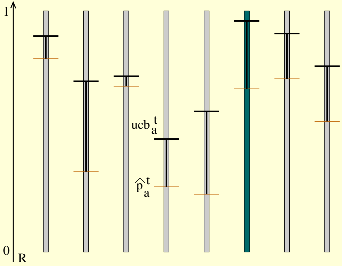
UCB (Auer *et al.*, 2002)

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



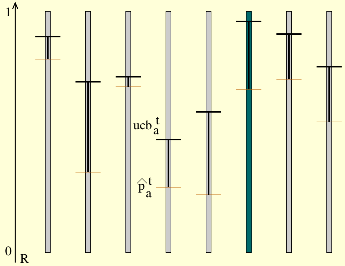
UCB (Auer *et al.*, 2002)

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



UCB (Auer *et al.*, 2002)

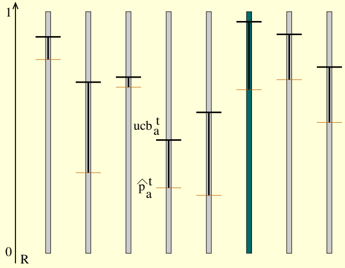
- Algorithm
  - Pull each arm once.
  - At time $t \in \{n, n+1, \dots\}$, for every arm a,
    $\text{ucb}_a^t \stackrel{\text{def}}{=} \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$; pull $\text{argmax}_a \, \text{ucb}_a^t$.



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

UCB (Auer *et al.*, 2002)

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



- 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.

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 some 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 some 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 some 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 some 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 some 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 some 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

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

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

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

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

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

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

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

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

Step 3: Bounding A

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

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

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

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

Step 4.1: Bounding B

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

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

1. We have:

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

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

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 $x \in \{\bar{u}_a^T, \bar{u}_a^T + 1, \dots, t\}$ 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!

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:

$$B \leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

Step 4.3: Bounding B

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

$$B \leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

Step 4.3: Bounding B

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

$$B \leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right)$$

Step 4.3: Bounding B

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

$$B \leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-4\ln(t)} + e^{-4\ln(t)}\right) \leq \sum_{t=0}^{T-1} t^2\left(\frac{2}{t^4}\right) \leq \sum_{t=0}^{\infty} \frac{2}{t^2} = \frac{\pi^2}{3}.$$

Step 4.3: Bounding B

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

$$B \leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_\star(y) + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(\mathbb{P}\left\{\hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}\right\} + \mathbb{P}\left\{\hat{p}_\star(y) < p_\star - \sqrt{\frac{2}{y}\ln(t)}\right\}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-4\ln(t)} + e^{-4\ln(t)}\right) \leq \sum_{t=0}^{T-1} t^2\left(\frac{2}{t^4}\right) \leq \sum_{t=0}^{\infty} \frac{2}{t^2} = \frac{\pi^2}{3}.$$

We are done!

Summary of Proof

- To upper-bound regret, 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.

Summary of Proof

- To upper-bound regret, 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 avoids the naive "union bound" to transform the random number pulls to a constant.

Multi-armed Bandits

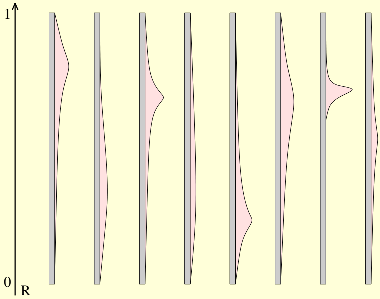
1. Concentration bounds
2. Analysis of UCB
3. Understanding Thompson Sampling
4. Other bandit problems

Thompson Sampling (Thompson, 1933)

- At time t, arm *a* has s_a^t successes (1's) and f_a^t failures (0's).

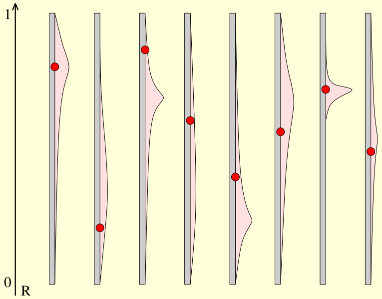
Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



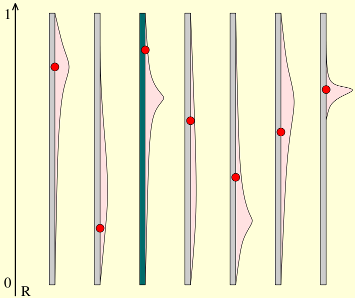
- Computational step: For every arm a, draw a sample

$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$

- Sampling step: Pull an arm a for which x_a^t is maximal.

Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



- Computational step: For every arm a, draw a sample

$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$

- Sampling step: Pull an arm a for which x_a^t is maximal.

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0 = \mathbb{P}\{w\}.$$

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0 = \mathbb{P}\{w\}.$$

- The process by which each w produces evidence e is known.
- Evidence samples $e_1, e_2, \ldots, e_m$ are produced i.i.d. by the unknown world w.

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0 = \mathbb{P}\{w\}.$$

- The process by which each w produces evidence e is known.
- Evidence samples $e_1, e_2, \ldots, e_m$ are produced i.i.d. by the unknown world w.
- How to continuously refine our belief distribution based on incoming evidence?

$$Belief_m = \mathbb{P}\{w|e_1, e_2, \ldots, e_m\}$$

Bayesian Inference

$$\begin{aligned}
Belief_{m+1}(w) &= \mathbb{P}\{w|e_1, e_2, \dots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_m|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_m, w\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{w|e_1, e_2, \dots, e_m\}\mathbb{P}\{e_1, e_2, \dots, e_m\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{Belief_m(w)\mathbb{P}\{e_{m+1}|w\}}{\sum_{w' \in W} Belief_m(w')\mathbb{P}\{e_{m+1}|w'\}}.
\end{aligned}$$

Bayesian Inference in Thompson Sampling

- For us w is the unknown bandit instance, but since the arms behave independently, we can think of each arm a's mean p_a as a world w, estimated based on its rewards (evidence).

Bayesian Inference in Thompson Sampling

- For us w is the unknown bandit instance, but since the arms behave independently, we can think of each arm a's mean p_a as a world w, estimated based on its rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.

Bayesian Inference in Thompson Sampling

- For us w is the unknown bandit instance, but since the arms behave independently, we can think of each arm a's mean p_a as a world w, estimated based on its rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

Bayesian Inference in Thompson Sampling

- For us w is the unknown bandit instance, but since the arms behave independently, we can think of each arm a's mean p_a as a world w, estimated based on its rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

- If e_{m+1} is a 0-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot (1 - x)}{\int_{y=0}^{1} Belief_m(y) \cdot (1 - y)}.$$

Bayesian Inference in Thompson Sampling

- For us w is the unknown bandit instance, but since the arms behave independently, we can think of each arm a's mean p_a as a world w, estimated based on its rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

- If e_{m+1} is a 0-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot (1 - x)}{\int_{y=0}^{1} Belief_m(y) \cdot (1 - y)}.$$

- We achieve exactly that by taking

$$Belief_m(x) = Beta_{s+1,f+1}(x)dx$$

when the first m pulls yield s 1's and f 0's!

Principle of Selecting Arm to Pull

- We have a belief distribution for each arm's mean.
- Together, these distributions represent a belief distribution over bandit instances.
- We sample a bandit instance I from the joint belief distribution, and
- We act optimally w.r.t. I.

Principle of Selecting Arm to Pull

- We have a belief distribution for each arm's mean.
- Together, these distributions represent a belief distribution over bandit instances.
- We sample a bandit instance I from the joint belief distribution, and
- We act optimally w.r.t. I.

- Alternative interpretation: the probability with which we pick an arm is our belief that it is optimal. For example, if $A = \{1, 2\}$, the probability of pulling 1 is $\mathbb{P}\{x_1^t > x_2^t\} =$

$$\int_{x_1=0}^{1} \int_{x_2=0}^{x_1} Beta_{s_1^t+1, f_1^t+1,}(x_1) Beta_{s_2^t+1, f_2^t+1,}(x_2) dx_2 dx_1.$$

Multi-armed Bandits

1. Concentration bounds
2. Analysis of UCB
3. Understanding Thompson Sampling
4. 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’re 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’re 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 aren’t 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.