CS 747, Autumn 2023: Lecture 4

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2023
Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- $\epsilon$-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm

- Understanding Thompson Sampling
- Concentration bounds

- Analysis of UCB
- Other bandit problems
Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- $\epsilon$-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm

Understanding Thompson Sampling
- Concentration bounds

Analysis of UCB
- Other bandit problems
Thompson Sampling (Thompson, 1933)

- At time $t$, arm $a$ has $s^t_a$ successes (1’s) and $f^t_a$ failures (0’s).
Thompson Sampling (Thompson, 1933)

- At time $t$, arm $a$ has $s^t_a$ successes (1’s) and $f^t_a$ failures (0’s).
- $Beta(s^t_a + 1, f^t_a + 1)$ 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 maximum.
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 maximum.
Bayesian Inference

- Bayes’ Rule of Probability for events $A$ and $B$:

$$
P(A|B) = \frac{P(B|A)P(A)}{P(B)}. $$

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 $0(w) = P\{w\}$.

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 refine our belief distribution based on incoming evidence?

Belief $m(w) = P\{w|e_1, e_2, \ldots, e_m\}.$
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 \textit{world} $w$ from among possible worlds $W$, in which we live.

- We maintain a \textit{belief} distribution over $w \in W$.

$$Belief_0(w) = \mathbb{P}\{w\}.$$
Bayesian Inference

Bayes’ Rule of Probability for events $A$ and $B$:

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

Application: there is an unknown *world* $w$ from among possible worlds $W$, in which we live.

We maintain a *belief* distribution over $w \in W$.

$$Belief_0(w) = P\{w\}.$$  

The process by/probability with 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 $B$:

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

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

\[
\text{Belief}_0(w) = P\{w\}.
\]

- The process by probability with 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 refine our belief distribution based on incoming evidence?

\[
\text{Belief}_m(w) = P\{w|e_1, e_2, \ldots, e_m\}.
\]
Bayesian Inference

\[ Belief_{m+1}(w) = \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \]
Bayesian Inference

\[
Belief_{m+1}(w) = \mathbb{P}\{w | e_1, e_2, \ldots, e_{m+1}\} = \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1} | w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\]
Bayesian Inference

\[ Belief_{m+1}(w) = \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1} \mid w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \]

\[ = \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m \mid w\}\mathbb{P}\{e_{m+1} \mid w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \]
Bayesian Inference

\[ \text{Belief}_{m+1}(w) = P\{w|e_1, e_2, \ldots, e_{m+1}\} \]

\[
= \frac{P\{e_1, e_2, \ldots, e_{m+1}|w\} P\{w\}}{P\{e_1, e_2, \ldots, e_{m+1}\}} \\
= \frac{P\{e_1, e_2, \ldots, e_m|w\} P\{e_{m+1}|w\} P\{w\}}{P\{e_1, e_2, \ldots, e_{m+1}\}} \\
= \frac{P\{e_1, e_2, \ldots, e_m, w\} P\{e_{m+1}|w\}}{P\{e_1, e_2, \ldots, e_{m+1}\}} \]
Bayesian Inference

\[
Belief_{m+1}(w) = \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \\
= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m, w\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
= \frac{\mathbb{P}\{w|e_1, e_2, \ldots, e_m\}\mathbb{P}\{e_1, e_2, \ldots, e_m\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\]
Bayesian Inference

\[ Belief_{m+1}(w) = \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \]

\[ = \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \]

\[ = \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m}|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \]

\[ = \frac{\mathbb{P}\{w|e_1, e_2, \ldots, e_m\}\mathbb{P}\{e_1, e_2, \ldots, e_{m}\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \]

\[ = \frac{Belief_m(w)\mathbb{P}\{e_{m+1}|w\}}{\sum_{w'\in W} Belief_m(w')\mathbb{P}\{e_{m+1}|w'\}}. \]
Bayesian Inference in Thompson Sampling

- View each arm $a$'s mean $p_a$ as world $w$, estimated from rewards (evidence).
Bayesian Inference in Thompson Sampling

- View each arm $a$’s mean $p_a$ as world $w$, estimated from rewards (evidence).
- $Belief_0$ over $p_a$ is typically set to $Uniform(0, 1)$, but need not.
Bayesian Inference in Thompson Sampling

- View each arm $a$'s mean $p_a$ as world $w$, estimated from 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 $x \in [0, 1]$

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

- We achieve exactly that by taking $Belief_m(x) = \text{Beta}(s+1, f+1)(x)$ when the first $m$ pulls yield $s$'s and $f$'0's!
Bayesian Inference in Thompson Sampling

- View each arm $a$'s mean $p_a$ as world $w$, estimated from 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 $x \in [0, 1]$

\[
Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_0^1 Belief_m(y) \cdot y}.
\]

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

\[
Belief_{m+1}(x) = \frac{Belief_m(x) \cdot (1 - x)}{\int_0^1 Belief_m(y) \cdot (1 - y)}.
\]
Bayesian Inference in Thompson Sampling

- View each arm $a$'s mean $p_a$ as world $w$, estimated from rewards (evidence).
- $\text{Belief}_0$ over $p_a$ is typically set to $\text{Uniform}(0, 1)$, but need not.
- If $e_{m+1}$ is a 1-reward, we must set for $x \in [0, 1]$
  \[
  \text{Belief}_{m+1}(x) = \frac{\text{Belief}_m(x) \cdot x}{\int_{y=0}^{1} \text{Belief}_m(y) \cdot y}.
  \]
- If $e_{m+1}$ is a 0-reward, we must set for $x \in [0, 1]$
  \[
  \text{Belief}_{m+1}(x) = \frac{\text{Belief}_m(x) \cdot (1 - x)}{\int_{y=0}^{1} \text{Belief}_m(y) \cdot (1 - y)}.
  \]
- We achieve exactly that by taking
  \[
  \text{Belief}_m(x) = \text{Beta}_{s+1, f+1}(x) dx
  \]
  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. $I$. 

Alternative view: 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 $P\{x_t \geq x_t^2\} = \int x_1^2 dx_1 = 0$. Beta $\alpha_{t+1} \geq x_1^2 = 0$ Beta $\beta_{t+1} \geq x_2^2 = 0$.
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 view: 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

$$\mathbb{P}\{x_1^t > x_2^t\} = \int_{x_1=0}^{1} \int_{x_2=0}^{x_1} \text{Beta}_{s_1^t+1, f_1^t+1}(x_1) \text{Beta}_{s_2^t+1, f_2^t+1}(x_2) dx_2 dx_1.$$
Multi-armed Bandits

1. Understanding Thompson Sampling

2. Concentration bounds
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

\[
\mathbb{P}\{\bar{x} \geq \mu + \epsilon\} \leq e^{-2u\epsilon^2}, \quad \text{and}
\]
\[
\mathbb{P}\{\bar{x} \leq \mu - \epsilon\} \leq e^{-2u\epsilon^2}.
\]
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 $\epsilon$, since $\bar{x} \in [0, 1]$. 

Shivaram Kalyanakrishnan (2023)
Applications

For given mistake probability $\delta$ and tolerance $\epsilon$, 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 $\mu$ by $\epsilon$ or more?

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

\[ P \{ \bar{x} \geq \mu + \epsilon \} \leq e^{-2u_0 \epsilon^2} \leq \delta. \]
Applications

For given mistake probability $\delta$ and tolerance $\epsilon$, 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 $\mu$ by $\epsilon$ or more?

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

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

For given mistake probability $\delta$ and tolerance $\epsilon$, 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 $\mu$ by $\epsilon$ or more?

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

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

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 $\mu$ by at most $\epsilon_0 = \underline{\text{__________}}$.  


Applications

For given mistake probability $\delta$ and tolerance $\epsilon$, 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 $\mu$ by $\epsilon$ or more? $u_0 = \lceil \frac{1}{2\epsilon^2} \ln(\frac{1}{\delta}) \rceil$ pulls are sufficient, since Hoeffding’s Inequality gives

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

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 $\mu$ by at most $\epsilon_0 = \underline{\underline{\_\_\_\_\_\_}}$.

We can write $\epsilon_0 = \sqrt{\frac{1}{2u} \ln(\frac{1}{\delta})}$; by Hoeffding’s Inequality:

$$P\{\bar{x} \geq \mu + \epsilon_0\} \leq e^{-2u(\epsilon_0)^2} \leq \delta.$$
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 $u; x_1, x_2, \ldots, x_u; \epsilon$ 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 $u; x_1, x_2, \ldots, x_u; \epsilon$ as defined earlier.

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

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

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

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

Shivaram Kalyanakrishnan (2023)
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

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

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

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

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

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

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

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

- Similar bounds can also be given when \( X \) has \textit{infinite support} (such as a Gaussian), but might need additional assumptions.
Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- $\epsilon$-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm

- Understanding Thompson Sampling
- Concentration bounds

- Analysis of UCB
- Other bandit problems