CS 747, Autumn 2022: Lecture 3

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

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

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

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

- Analysis of UCB
- Understanding Thompson Sampling
- Other bandit problems

Multi-armed Bandits

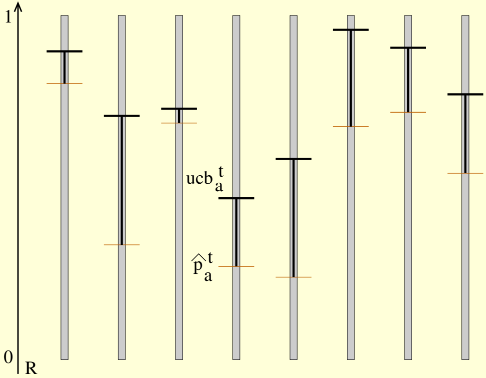
- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- ϵ-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret

- UCB, KL-UCB algorithms
- Thompson Sampling algorithm
- Concentration bounds

- Analysis of UCB
- Understanding Thompson Sampling
- Other bandit problems

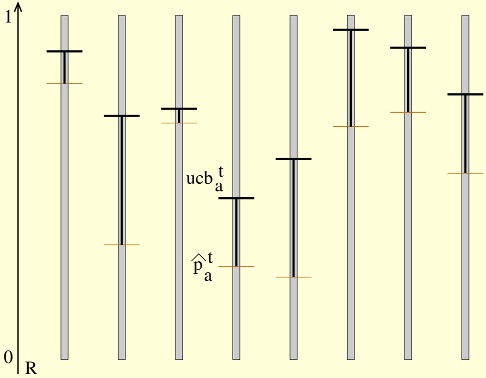
Upper Confidence Bounds = UCB (Auer et al., 2002)

- At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
- $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
- u_a^t the number of times a has been sampled at time t.



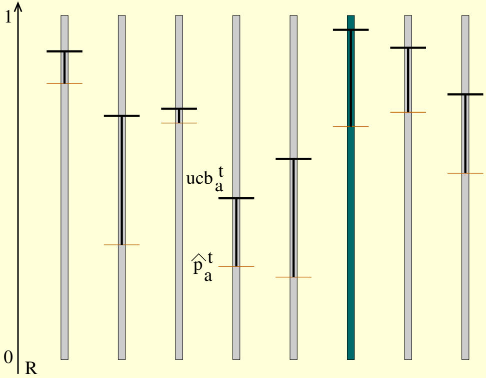
Upper Confidence Bounds = UCB (Auer et al., 2002)

- At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
- $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
- u_a^t the number of times a has been sampled at time t.
- Pull an arm a for which ucb_a^t is maximum.



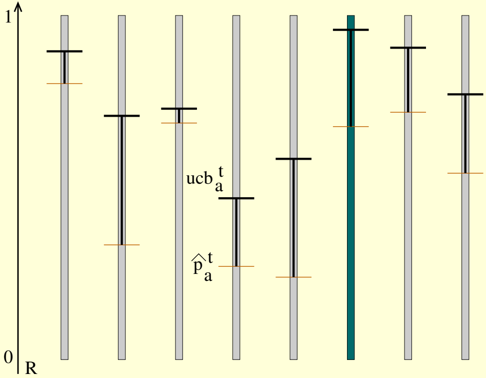
Upper Confidence Bounds = UCB (Auer et al., 2002)

- At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
- $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
- u_a^t the number of times a has been sampled at time t.
- Pull an arm a for which ucb_a^t is maximum.



Upper Confidence Bounds = UCB (Auer et al., 2002)

- At time t, for every arm a, define $\text{ucb}_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$.
- $\hat{p}_a^t$ is the empirical mean of rewards from arm a.
- u_a^t the number of times a has been sampled at time t.
- Pull an arm a for which ucb_a^t is maximum.



Achieves regret of $O(\log(T))$: optimal dependence on T.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c \ln(\ln(t))\}$, where $c \geq 3$.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}^t_a = \max\{q \in [\hat{p}^t_a, 1] \text{ s. t. } u^t_a KL(\hat{p}^t_a, q) \leq \ln(t) + c \ln(\ln(t))\}$, where $c \geq 3$.

  Equivalently, ucb-kl^t_a is the solution $q \in [\hat{p}^t_a, 1]$ to $KL(\hat{p}^t_a, q) = \frac{\ln(t) + c\ln(\ln(t))}{u^t_a}$.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c \ln(\ln(t))\}$, where $c \geq 3$.

  Equivalently, ucb-kl_a^t is the solution $q \in [\hat{p}_a^t, 1]$ to $KL(\hat{p}_a^t, q) = \frac{\ln(t) + c\ln(\ln(t))}{u_a^t}$.

  KL-UCB algorithm: at step t, pull $\operatorname{argmax}_{a \in A} \text{ucb-kl}_a^t$.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\}$, where $c \geq 3$.

  Equivalently, ucb-kl_a^t is the solution $q \in [\hat{p}_a^t, 1]$ to $KL(\hat{p}_a^t, q) = \frac{\ln(t)+c\ln(\ln(t))}{u_a^t}$.

  KL-UCB algorithm: at step t, pull $\operatorname{argmax}_{a \in A} \text{ucb-kl}_a^t$.

- Observe that $KL(\hat{p}_a^t, q)$ monotonically increases with q, and
  - $KL(\hat{p}_a^t, \hat{p}_a^t) = 0$;
  - $KL(\hat{p}_a^t, 1) = \infty$.

  Easy to compute ucb-kl_a^t numerically (for example through binary search).

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\}$, where $c \geq 3$.

  Equivalently, ucb-kl_a^t is the solution $q \in [\hat{p}_a^t, 1]$ to $KL(\hat{p}_a^t, q) = \frac{\ln(t) + c\ln(\ln(t))}{u_a^t}$.

  KL-UCB algorithm: at step t, pull $\operatorname{argmax}_{a \in A} \text{ucb-kl}_a^t$.

- Observe that $KL(\hat{p}_a^t, q)$ monotonically increases with q, and
  - $KL(\hat{p}_a^t, \hat{p}_a^t) = 0$;
  - $KL(\hat{p}_a^t, 1) = \infty$.

  Easy to compute ucb-kl_a^t numerically (for example through binary search).

- ucb-kl_a^t is a tighter confidence bound than ucb_a^t.

KL-UCB (Garivier and Cappé, 2011)

- Identical to UCB algorithm on previous slide, except for a different definition of the upper confidence bound.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c \ln(\ln(t))\}$, where $c \geq 3$.

  Equivalently, ucb-kl_a^t is the solution $q \in [\hat{p}_a^t, 1]$ to $KL(\hat{p}_a^t, q) = \frac{\ln(t) + c\ln(\ln(t))}{u_a^t}$.

  KL-UCB algorithm: at step t, pull $\operatorname{argmax}_{a \in A} \text{ucb-kl}_a^t$.

- Observe that $KL(\hat{p}_a^t, q)$ monotonically increases with q, and
  - $KL(\hat{p}_a^t, \hat{p}_a^t) = 0$;
  - $KL(\hat{p}_a^t, 1) = \infty$.

  Easy to compute ucb-kl_a^t numerically (for example through binary search).

- ucb-kl_a^t is a tighter confidence bound than ucb_a^t.

  Regret of KL-UCB asymptotically matches Lai and Robbins' lower bound!

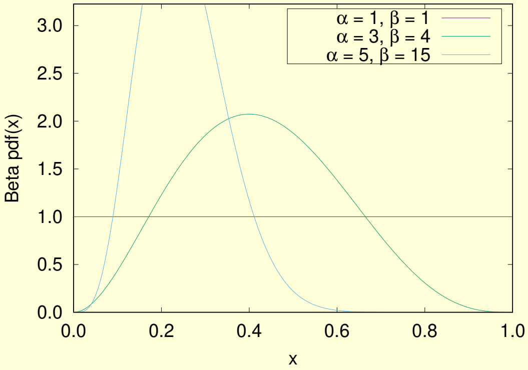
Multi-armed Bandits

1. UCB, KL-UCB algorithms
2. Thompson Sampling algorithm
3. Concentration bounds

Background: Beta Distribution

- Beta(α, β) defined on $[0, 1]$. Two parameters: α and β.

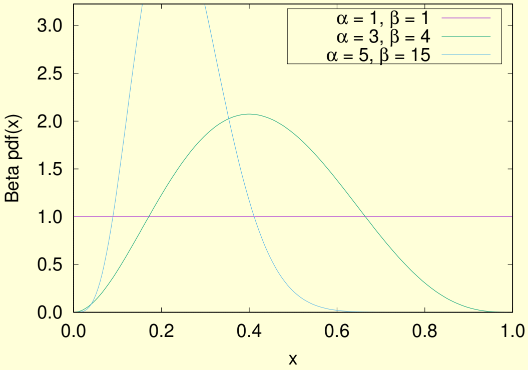
$$\text{Mean} = \frac{\alpha}{\alpha + \beta}; \quad \text{Variance} = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

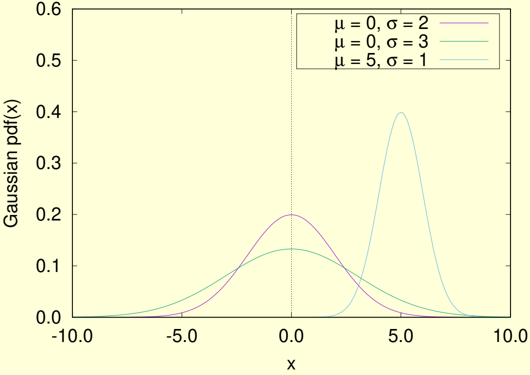


Background: Beta Distribution

- Beta(α, β) defined on $[0, 1]$. Two parameters: α and β.

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}; \quad \text{Variance} = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$



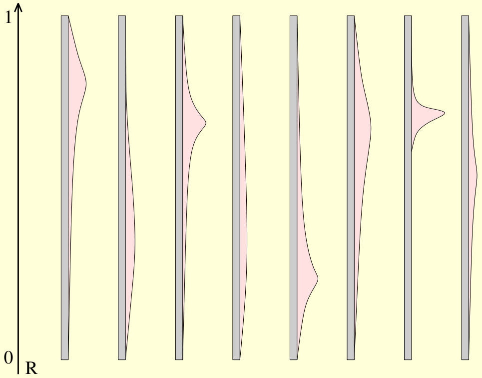


Thompson Sampling (Thompson, 1933)

- At time t, let arm *a* have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).

Thompson Sampling (Thompson, 1933)

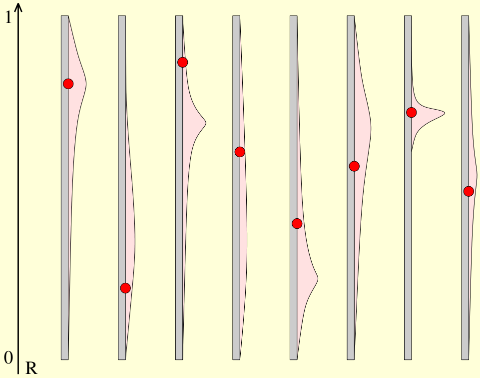
- At time t, let arm *a* have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm *a*.
- Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.



Thompson Sampling (Thompson, 1933)

- At time t, let arm a have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
- Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.

- Computational step: For every arm a, draw a sample (in agent's mind)
$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$
- Sampling step: Pull (in real world) arm a for which x_a^t is maximum.



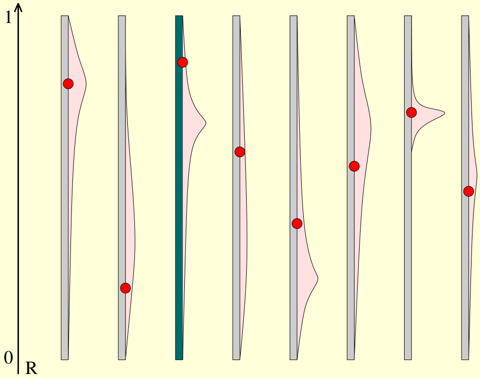
Thompson Sampling (Thompson, 1933)

- At time t, let arm a have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
- Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.

- Computational step: For every arm a, draw a sample (in agent's mind)

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

- Sampling step: Pull (in real world) arm a for which x_a^t is maximum.

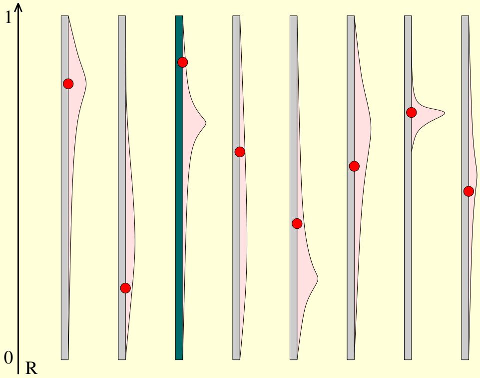


Thompson Sampling (Thompson, 1933)

- At time t, let arm a have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
- Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$.

- Computational step: For every arm a, draw a sample (in agent's mind)
$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$
- Sampling step: Pull (in real world) arm a for which x_a^t is maximum.

Achieves optimal regret (Kaufmann et al., 2012); is excellent in practice (Chapelle and Li, 2011).



Multi-armed Bandits

1. UCB, KL-UCB algorithms
2. Thompson Sampling algorithm
3. 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

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

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 $x_1, x_2, \ldots, x_u$ 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. UCB, KL-UCB algorithms
2. Thompson Sampling algorithm
3. Concentration bounds