CS 747, Autumn 2023: Lecture 3

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
- 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

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

A Lower Bound on Regret

Paraphrasing Lai and Robbins (1985; see Theorem 2).

Let L be an algorithm such that for every bandit instance $I \in \bar{\mathcal{I}}$ and for every $\alpha > 0$, as $T \to \infty$:

$$R_T(L, I) = o(T^\alpha).$$

Then, for every bandit instance $I \in \bar{\mathcal{I}}$, as $T \to \infty$:

$$\frac{R_T(L, I)}{\ln(T)} \geq \sum_{a: p_a(I) \neq p^\star(I)} \frac{p^\star(I) - p_a(I)}{KL(p_a(I), p^\star(I))},$$

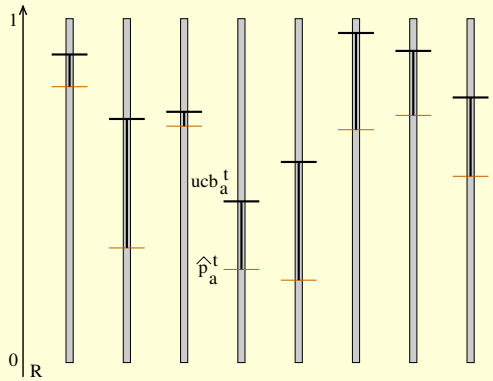
where for $x, y \in [0, 1)$, $KL(x, y) \stackrel{\text{def}}{=} x \ln \frac{x}{y} + (1 - x) \ln \frac{1-x}{1-y}$.

Multi-armed Bandits

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

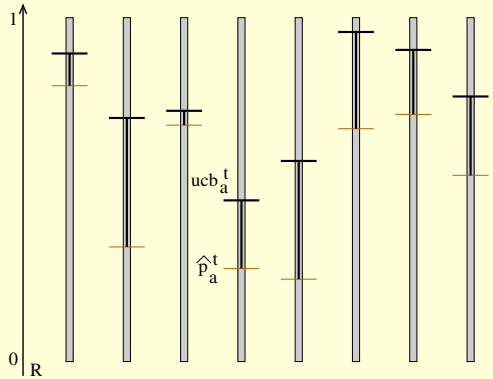
Upper Confidence Bounds = UCB (Auer et al., 2002)

- 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.



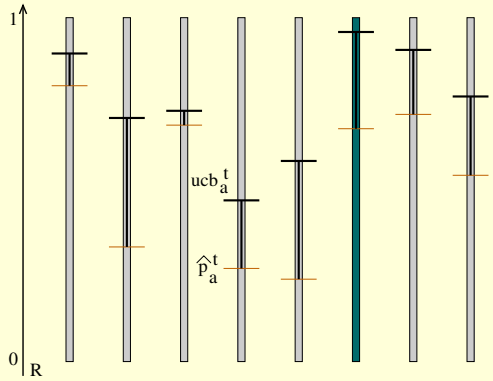
Upper Confidence Bounds = UCB (Auer et al., 2002)

- 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.



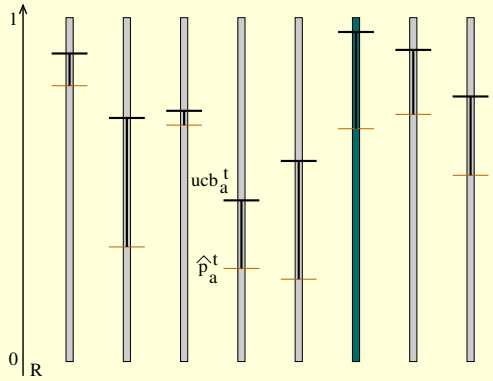
Upper Confidence Bounds = UCB (Auer et al., 2002)

- 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
up to a constant factor.

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.

  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.

  $\text{ucb-kl}_a^t = \max\{q \in [\hat{p}_a^t, 1] \text{ s. t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\}$, 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 (Garivier and Cappé, 2011)

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

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

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

- 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.
  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 $\text{argmax}_{a \in A}$ 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!

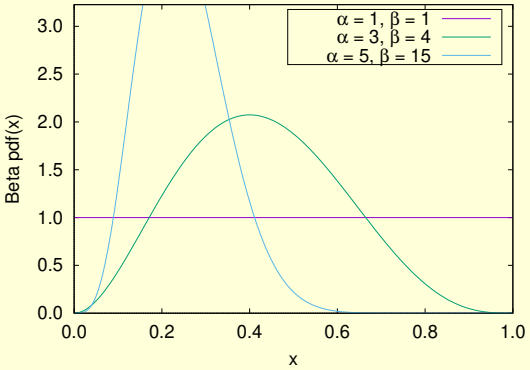
Multi-armed Bandits

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

Background: Beta Distribution

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

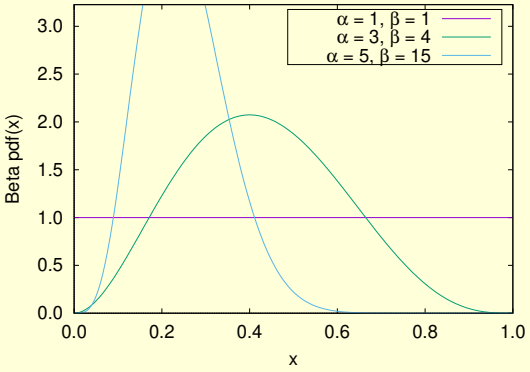
$$\text{Mean} = \frac{\alpha}{\alpha + \beta}; \quad \text{Variance} = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

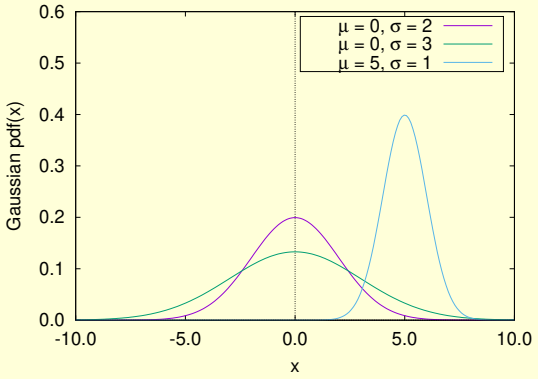


Background: Beta Distribution

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

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



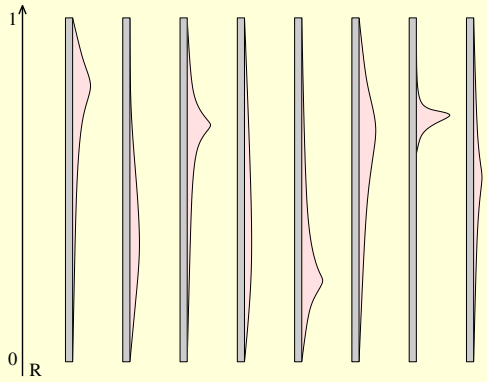


Thompson Sampling (Thompson, 1933)

- 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)

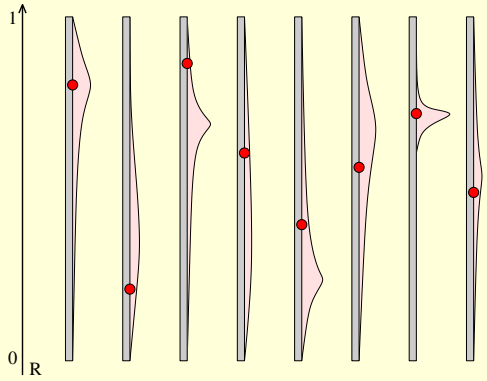
- At time t, let arm a have s_a^t successes (1's/heads) and f_a^t failures (0's/tails).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about the true mean of arm a.
- Mean = $\frac{s_a^t+1}{s_a^t+f_a^t+2}$; variance = $\frac{\left(s_a^t+1\right)\left(f_a^t+1\right)}{\left(s_a^t+f_a^t+2\right)^2\left(s_a^t+f_a^t+3\right)}$.



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 = $\frac{\left(s_a^t+1\right)\left(f_a^t+1\right)}{\left(s_a^t+f_a^t+2\right)^2\left(s_a^t+f_a^t+3\right)}$.

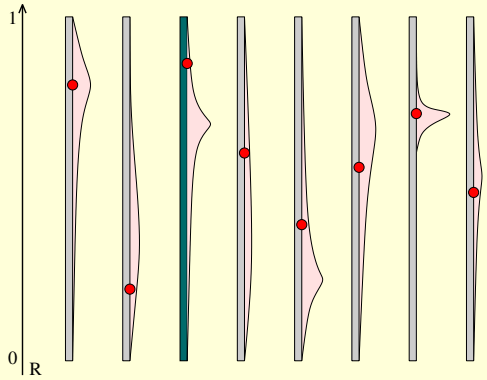
- Computational step: For every arm a, draw a sample (in agent's mind)
$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$
- Sampling step: Pull (in real world) arm a for which x_a^t is maximum.



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 = $\frac{\left(s_a^t+1\right)\left(f_a^t+1\right)}{\left(s_a^t+f_a^t+2\right)^2\left(s_a^t+f_a^t+3\right)}$.

- 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.

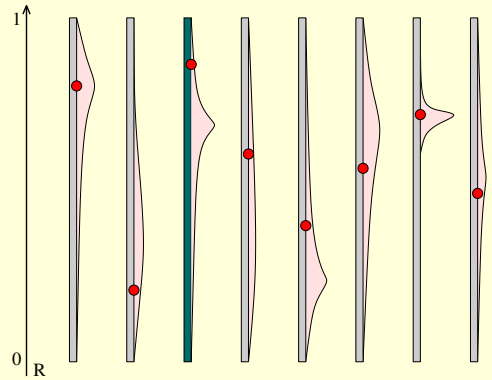


Thompson Sampling (Thompson, 1933)

- 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 = $\frac{\left(s_a^t+1\right)\left(f_a^t+1\right)}{\left(s_a^t+f_a^t+2\right)^2\left(s_a^t+f_a^t+3\right)}$.

- 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

- 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