Lecture 2 - Upper Confidence Bound Algorithm for Bandits

Pratik Gajane

September 12, 2022

2AMM20 Research Topics in Data Mining
Eindhoven University of Technology

A Quick Recap of Lecture 1

- Introduction to reinforcement learning.
- Mathematical formulation of a reinforcement learning problem.
- Formulating RL with multi-armed bandits and its variants.
- Formulating RL with Markov decision processes.

Lecture 2 : Outline

- Introduction to Bandits and Mathematical Setting
- Greedy : A Simple Solution (and why it does not work?)
- Acting optimistically : Upper Confidence Bound algorithm.

Introduction

What's in a Name? Why "Bandits"?

A single-armed bandit.

One arm $\equiv$ one choice.

What's in a Name? Why "Bandits"?

A single-armed bandit.
One arm ≡ one choice.

A multi-armed bandit.
Multiple arms ≡ multiple choices.

Multi-Armed Bandits

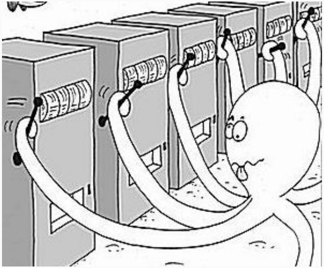


Image source : *Microsoft research*

- Agent faces repeated choice among K different actions/arms.

Multi-Armed Bandits

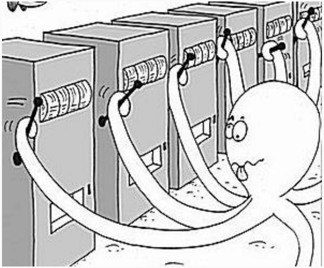


Image source : *Microsoft research*

- Agent faces repeated choice among K different actions/arms.
- Agent acts according to some policy π.
- At each time step t, the agent selects an action and then receives a numerical reward for that action.

Multi-Armed Bandits

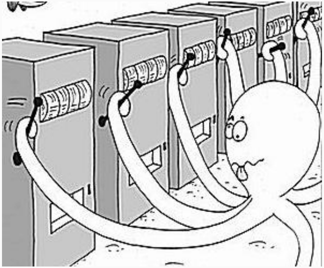


Image source : *Microsoft research*

- Agent faces repeated choice among K different actions/arms.
- Agent acts according to some policy π.
- At each time step t, the agent selects an action and then receives a numerical reward for that action.(Bandit feedback).

Multi-Armed Bandits

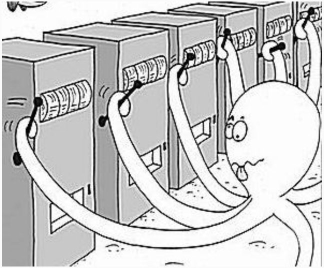


Image source : *Microsoft research*

- Agent faces repeated choice among K different actions/arms.
- Agent acts according to some policy π.
- At each time step t, the agent selects an action and then receives a numerical reward for that action.(Bandit feedback).
- Agent learns only through received rewards. No other way to learn.

Multi-Armed Bandits

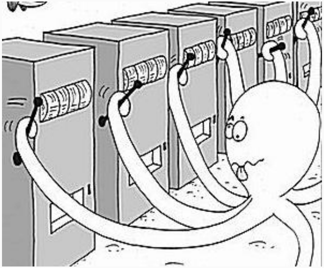


Image source : *Microsoft research*

- Agent faces repeated choice among K different actions/arms.
- Agent acts according to some policy π.
- At each time step t, the agent selects an action and then receives a numerical reward for that action.(Bandit feedback).
- Agent learns only through received rewards. No other way to learn.
- Goal : Maximize the sum of received rewards.

Exploration/Exploitation Dilemma

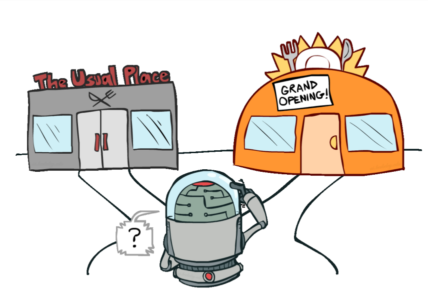


Image source: *UC Berkeley AI course, lecture 11*

- Exploit. Choose actions tried in the past and found to be rewarding.

Exploration/Exploitation Dilemma

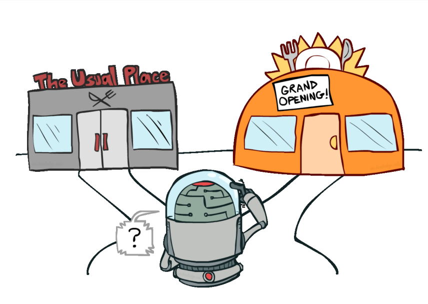


Image source: *UC Berkeley AI course, lecture 11*

- Exploit. Choose actions tried in the past and found to be rewarding.
- Explore. Choose unexplored actions to see if they are more rewarding.

Exploration/Exploitation Dilemma

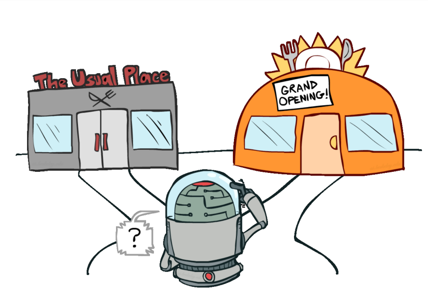


Image source: *UC Berkeley AI course, lecture 11*

- Exploit. Choose actions tried in the past and found to be rewarding.
- Explore. Choose unexplored actions to see if they are more rewarding.
- Neither exploration nor exploitation can be pursued exclusively.

Exploration/Exploitation Dilemma

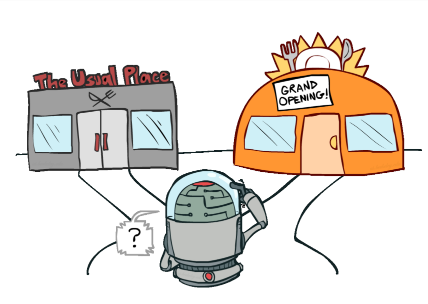


Image source: *UC Berkeley AI course, lecture 11*

- Exploit. Choose actions tried in the past and found to be rewarding.
- Explore. Choose unexplored actions to see if they are more rewarding.
- Neither exploration nor exploitation can be pursued exclusively.
- A good solution balances exploration and exploitation.

Applications!

- Clinical trials
- Recommendation systems
- Ad placement
- Dynamic pricing
- And many more . . .

Mathematical setting

Stationary Stochastic Bandits

- Number of arms $= K$.

Stationary Stochastic Bandits

- Number of arms = K.
- Reward for arm $a \sim X_a$ with mean μ_a.
- $X_1, X_2, \ldots X_K$ are unknown stationary distributions.

Stationary Stochastic Bandits

- Number of arms $= K$.
- Reward for arm $a \sim X_a$ with mean μ_a.
- $X_1, X_2, \dots X_K$ are unknown stationary distributions.
- At each time step $t = 1, \dots, T$, the agent,
    - acts according to a policy π and chooses an arm $a(t)$, and

Stationary Stochastic Bandits

- Number of arms $= K$.
- Reward for arm $a \sim X_a$ with mean μ_a.
- $X_1, X_2, \dots X_K$ are unknown stationary distributions.
- At each time step $t = 1, \dots, T$, the agent,
  - acts according to a policy π and chooses an arm $a(t)$, and
  - receives a numerical reward $r(t) \sim X_{a(t)}$.

Stationary Stochastic Bandits

- Number of arms $= K$.
- Reward for arm $a \sim X_a$ with mean μ_a.
- $X_1, X_2, \ldots X_K$ are unknown stationary distributions.
- At each time step $t = 1, \ldots, T$, the agent,
  - acts according to a policy π and chooses an arm $a(t)$, and
  - receives a numerical reward $r(t) \sim X_{a(t)}$.
- T is called the horizon.

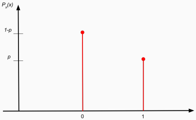
Distributional Assumptions?

Distributions $X_1, \ldots, X_K$ are unknown to the agent, but we may make some assumptions. E.g.,

Distributional Assumptions?

Distributions $X_1, \ldots, X_K$ are unknown to the agent, but we may make some assumptions. E.g.,

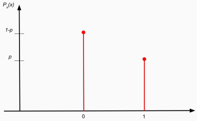
- X_a is Bernoulli with unknown mean $\mu_a \in [0, 1]$.



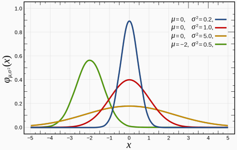
Distributional Assumptions?

Distributions $X_1, \ldots, X_K$ are unknown to the agent, but we may make some assumptions. E.g.,

- X_a is Bernoulli with unknown mean $\mu_a \in [0, 1]$.



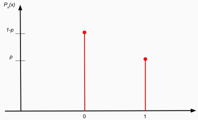
- X_a is Gaussian with unit variance and unknown mean $\mu_a \in \mathbb{R}$.



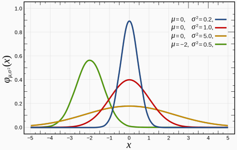
Distributional Assumptions?

Distributions $X_1, \ldots, X_K$ are unknown to the agent, but we may make some assumptions. E.g.,

- X_a is Bernoulli with unknown mean $\mu_a \in [0, 1]$.



- X_a is Gaussian with unit variance and unknown mean $\mu_a \in \mathbb{R}$.



Which assumption do we make? We will see in due time.

Stationary Stochastic Bandits : Example

a_1, Bernoulli, mean $\mu_1 = 0.9$ a_2, Bernoulli, mean $\mu_2 = 0.8$

- Number of arms $= K = 2$.

Stationary Stochastic Bandits : Example

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Number of arms $= K = 2$.
- Reward for arm $a_1 \sim$ Bernoulli with mean $\mu_1 = 0.9$.
  Reward for arm $a_2 \sim$ Bernoulli with mean $\mu_2 = 0.8$

Stationary Stochastic Bandits : Example

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Number of arms $= K = 2$.
- Reward for arm $a_1 \sim$ Bernoulli with mean $\mu_1 = 0.9$.
  Reward for arm $a_2 \sim$ Bernoulli with mean $\mu_2 = 0.8$
- Agent policy π : Choose arms alternatingly.

Stationary Stochastic Bandits: Example

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Number of arms $= K = 2$.
- Reward for arm $a_1 \sim$ Bernoulli with mean $\mu_1 = 0.9$.
  Reward for arm $a_2 \sim$ Bernoulli with mean $\mu_2 = 0.8$
- Agent policy π: Choose arms alternatingly.
- The agent,
  - at $t = 1, 3, \ldots$, picks arm a_1, reward $r(t) \sim$ Bernoulli with $\mu_1 = 0.9$;
  - at $t = 2, 4, \ldots$, picks arm a_2, reward $r(t) \sim$ Bernoulli with $\mu_2 = 0.8$.

Random Variables, Expectation and Indicator Function

- Random variable : A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.

Random Variables, Expectation and Indicator Function

- Random variable : A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.
- Expectation of a random variable $x = \mathbb{E}[x] := \sum_i i \cdot \mathbb{P}(x = i)$.

Random Variables, Expectation and Indicator Function

- Random variable : A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.
- Expectation of a random variable $x = \mathbb{E}[x] := \sum_i i \cdot \mathbb{P}(x = i)$.
- Expected value of a random variable is the mean of the related stochastic process.

Random Variables, Expectation and Indicator Function

- Random variable : A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.
- Expectation of a random variable $x = \mathbb{E}[x] \coloneqq \sum_i i \cdot \mathbb{P}(x = i)$.
- Expected value of a random variable is the mean of the related stochastic process.
- Expectation is linear i.e.,

$$\mathbb{E}[x_1 + x_2 + \cdots + x_n] = \mathbb{E}[x_1] + \mathbb{E}[x_2] + \cdots + \mathbb{E}[x_n].$$

Random Variables, Expectation and Indicator Function

- Random variable : A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.
- Expectation of a random variable $x = \mathbb{E}[x] := \sum_i i \cdot \mathbb{P}(x = i)$.
- Expected value of a random variable is the mean of the related stochastic process.
- Expectation is linear i.e.,

$$\mathbb{E}[x_1 + x_2 + \cdots + x_n] = \mathbb{E}[x_1] + \mathbb{E}[x_2] + \cdots + \mathbb{E}[x_n].$$

- Indicator function $\mathbb{I}(E) = \begin{cases} 1, & \text{if } E \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$

Random Variables, Expectation and Indicator Function

- Random variable: A quantity which depends on a random/stochastic process.
  e.g., outcome of a coin toss, reward drawn from a stochastic distribution.
- Expectation of a random variable $x = \mathbb{E}[x] := \sum_i i \cdot \mathbb{P}(x = i)$.
- Expected value of a random variable is the mean of the related stochastic process.
- Expectation is linear i.e.,

$$\mathbb{E}[x_1 + x_2 + \cdots + x_n] = \mathbb{E}[x_1] + \mathbb{E}[x_2] + \cdots + \mathbb{E}[x_n].$$

- Indicator function $\mathbb{I}(E) = \begin{cases} 1, & \text{if } E \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$

  e.g., count of occurrences of $E = \sum_{t=1}^{T} \mathbb{I}(E)$.

Optimal Policy

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Goal : Maximize expected cumulative reward.
- Expected cumulative reward of policy π till $T := \mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]$.

Optimal Policy

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Goal : Maximize expected cumulative reward.
- Expected cumulative reward of policy π till $T := \mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]$.
- Optimal policy $\pi_* := \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]$.

Optimal Policy

a_1, Bernoulli, mean $\mu_1 = 0.9$

a_2, Bernoulli, mean $\mu_2 = 0.8$

- Goal : Maximize expected cumulative reward.
- Expected cumulative reward of policy π till $T := \mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]$.
- Optimal policy $\pi_* := \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]$.
- Policy π_* : Play the optimal arm with mean reward $\mu_* := \max_a \mu_a$.

Performance Measure : Regret

- If the agent acts according to policy π_* at t, then
  it receives the optimal expected reward $\mu_* := \max_a \mu_a$.

Performance Measure : Regret

- If the agent acts according to policy π_* at t, then
  it receives the optimal expected reward $\mu_* := \max_a \mu_a$.
- If the agent acts according to policy π_* from $t = 1, \ldots, T$, then
  it receives the optimal expected cumulative reward $= T\mu_*$.

Performance Measure : Regret

- If the agent acts according to policy π_* at t, then
  it receives the optimal expected reward $\mu_* := \max_a \mu_a$.
- If the agent acts according to policy π_* from $t = 1, \ldots, T$, then
  it receives the optimal expected cumulative reward $= T\mu_*$.
- *Regret is a measure of the total mistake cost.*
  How far is the agent's performance from the optimal performance?

Performance Measure : Regret

- If the agent acts according to policy π_* at t, then
  it receives the optimal expected reward $\mu_* := \max_a \mu_a$.
- If the agent acts according to policy π_* from $t = 1, \ldots, T$, then
  it receives the optimal expected cumulative reward $= T\mu_*$.
- *Regret is a measure of the total mistake cost.*
  How far is the agent's performance from the optimal performance?
- Regret $= \mathfrak{R}_\pi(T) := \underbrace{T\mu_*}_{\text{Optimal expected cumulative reward}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]}_{\text{Expected cumulative reward of } \pi}$

Performance Measure : Regret

- If the agent acts according to policy π_* at t, then
  it receives the optimal expected reward $\mu_* := \max_a \mu_a$.
- If the agent acts according to policy π_* from $t = 1, \ldots, T$, then
  it receives the optimal expected cumulative reward $= T\mu_*$.
- *Regret is a measure of the total mistake cost.*
  How far is the agent's performance from the optimal performance?
- Regret $= \mathfrak{R}_\pi(T) := \underbrace{T\mu_*}_{\text{Optimal expected cumulative reward}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]}_{\text{Expected cumulative reward of } \pi}$
- Minimizing regret $\equiv$ Maximizing expected cumulative reward.

Decomposing Regret into Arms I

Suboptimality gap $\Delta_a \coloneqq \mu_* - \mu_a$

Decomposing Regret into Arms I

Suboptimality gap $\Delta_a := \mu_* - \mu_a$

$N_a(T) :=$ Number of times arm a is played till T

Decomposing Regret into Arms I

Suboptimality gap $\Delta_a := \mu_* - \mu_a$

$$
\begin{aligned}
N_a(T) &:= \text{ Number of times arm } a \text{ is played till } T \\
&= \sum_{t=1}^{T} \mathbb{I}(a(t) = a) \qquad \text{where } a(t) \text{ is the arm selected at time } t.
\end{aligned}
$$

Decomposing Regret into Arms I

Suboptimality gap $\Delta_a := \mu_* - \mu_a$

$$N_a(T) := \text{ Number of times arm } a \text{ is played till } T$$

$$= \sum_{t=1}^{T} \mathbb{I}(a(t) = a) \qquad \text{where } a(t) \text{ is the arm selected at time } t.$$

Lemma

$$Regret = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

Decomposing Regret into Arms II

Lemma

$$\textit{Regret} = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

Decomposing Regret into Arms II

Lemma

$$Regret = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

Decomposing Regret into Arms II

Lemma

$$\mathit{Regret} = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] = \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

Decomposing Regret into Arms II

Lemma

$$Regret = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\begin{aligned}
\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] &= \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right]
\end{aligned}$$

Decomposing Regret into Arms II

Lemma

$$Regret = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] = \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{a(t)}\right]$$

Decomposing Regret into Arms II

Lemma

$$\mathit{Regret} = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\begin{aligned}
\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] &= \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{a(t)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{I}(a(t) = a)\Delta_a\right]
\end{aligned}$$

Decomposing Regret into Arms II

Lemma

$$Regret = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\begin{aligned}
\mathfrak{R}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] &= \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{a(t)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{I}(a(t) = a)\Delta_a\right] = \mathbb{E}\left[\sum_{a=1}^{K} \Delta_a \sum_{t=1}^{T} \mathbb{I}(a(t) = a)\right]
\end{aligned}$$

Decomposing Regret into Arms II

Lemma

$$\mathit{Regret} = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\begin{aligned}
\mathfrak{R}(T) &= T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] = \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{a(t)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{I}(a(t) = a)\Delta_a\right] = \mathbb{E}\left[\sum_{a=1}^{K} \Delta_a \sum_{t=1}^{T} \mathbb{I}(a(t) = a)\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \Delta_a N_a(T)\right]
\end{aligned}$$

Decomposing Regret into Arms II

Lemma

$$\textit{Regret} = \mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

$$\begin{aligned}
\mathfrak{R}(T) &= T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] = \sum_{t=1}^{T} \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} [\mu_* - r(t)]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{a(t)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{I}(a(t) = a)\Delta_a\right] = \mathbb{E}\left[\sum_{a=1}^{K} \Delta_a \sum_{t=1}^{T} \mathbb{I}(a(t) = a)\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \Delta_a N_a(T)\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].
\end{aligned}$$

Target Regret?

- Two arms with Bernoulli rewards, $\mu_1 = 0.9$ and $\mu_2 = 0.8$.

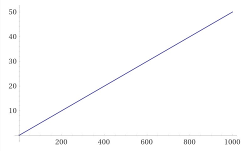
Target Regret?

- Two arms with Bernoulli rewards, $\mu_1 = 0.9$ and $\mu_2 = 0.8$.
- Policy π : Play each arm with probability 0.5.

Target Regret?

- Two arms with Bernoulli rewards, $\mu_1 = 0.9$ and $\mu_2 = 0.8$.
- Policy π : Play each arm with probability 0.5.

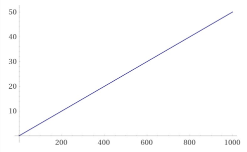
$$\begin{aligned}\text{Regret of } \pi &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\ &= (0.9 - 0.9)\tfrac{T}{2} + (0.9 - 0.8)\tfrac{T}{2} \\ &= 0.05T \quad (\text{linear regret!}).\end{aligned}$$



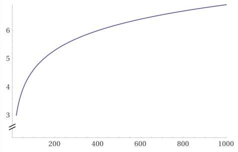
Target Regret?

- Two arms with Bernoulli rewards, $\mu_1 = 0.9$ and $\mu_2 = 0.8$.
- Policy π : Play each arm with probability 0.5.

$$\begin{aligned}\text{Regret of } \pi &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\ &= (0.9 - 0.9)\tfrac{T}{2} + (0.9 - 0.8)\tfrac{T}{2} \\ &= 0.05T \quad (\text{linear regret!}).\end{aligned}$$



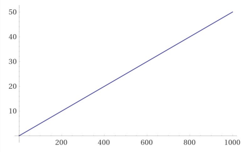
- A policy with sub-linear regret is said to be learning.



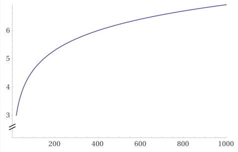
Target Regret?

- Two arms with Bernoulli rewards, $\mu_1 = 0.9$ and $\mu_2 = 0.8$.
- Policy π : Play each arm with probability 0.5.

$$\begin{aligned}\text{Regret of } \pi &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\ &= (0.9 - 0.9)\tfrac{T}{2} + (0.9 - 0.8)\tfrac{T}{2} \\ &= 0.05T \quad (\text{linear regret!}).\end{aligned}$$



- A policy with sub-linear regret is said to be learning.
- **Goal : Construct an algorithm with sub-linear regret**.



Solutions

How to Minimize Regret?

- Suboptimality gap $\Delta_a := \mu_* - \mu_a$.
- $N_a(T) :=$ Number of times arm a is played till $T = \sum_{t=1}^{T} \mathbb{I}(a(t) = a)$.
- Regret $\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$.

How to Minimize Regret?

- Suboptimality gap $\Delta_a := \mu_* - \mu_a$.
- $N_a(T) :=$ Number of times arm a is played till $T = \sum_{t=1}^{T} \mathbb{I}(a(t) = a)$.
- Regret $\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$.
- For large gaps Δ_a, keep count $N_a(T)$ small.

How to Minimize Regret?

- Suboptimality gap $\Delta_a := \mu_* - \mu_a$.
- $N_a(T) :=$ Number of times arm a is played till $T = \sum_{t=1}^{T} \mathbb{I}(a(t) = a)$.
- Regret $\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$.
- For large gaps Δ_a, keep count $N_a(T)$ small.
- If mean reward μ's are known, simply pick the arm with $\mu_* = \arg\max_a \mu_a$.
  But they are unknown. So, build an estimate $\hat{\mu}$.

How to Minimize Regret?

- Suboptimality gap $\Delta_a := \mu_* - \mu_a$.
- $N_a(T) :=$ Number of times arm a is played till $T = \sum_{t=1}^{T} \mathbb{I}(a(t) = a)$.
- Regret $\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$.
- For large gaps Δ_a, keep count $N_a(T)$ small.
- If mean reward μ's are known, simply pick the arm with $\mu_* = \arg\max_a \mu_a$.
  But they are unknown. So, build an estimate $\hat{\mu}$.
- $\hat{\mu}_a(t) =$ Empirical mean of arm a at time t
  $=$ Average of the received rewards from arm a till t
  $= \frac{1}{N_a(t)} \sum_{\tau=1}^{t} (r(\tau) | a(\tau) = a)$.

Greedy Algorithm

$$\text{Greedy}: \begin{array}{l}\text{Choose each action once.} \\ \text{Then choose the action with the highest empirical mean.}\end{array}$$

Greedy Algorithm

Greedy : Choose each action once.
Then choose the action with the highest empirical mean.

Algorithm Greedy algorithm

for $t = 1, \dots, K$ **do**
    Choose each arm once.
end for
for $t = K + 1, \dots$ **do**
    Compute empirical means $\hat{\mu}_1(t-1), \dots, \hat{\mu}_K(t-1)$.
    Select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.
end for

Greedy Algorithm

Greedy : Choose each action once.
Then choose the action with the highest empirical mean.

Algorithm Greedy algorithm

```
for t = 1, ..., K do
    Choose each arm once.
end for
for t = K + 1, ... do
    Compute empirical means μ̂_1(t − 1), ..., μ̂_K(t − 1).
    Select arm a(t) = arg max_a μ̂_a(t − 1).
end for
```

Greedy algorithm has linear regret!

Why Does Greedy Fail?

Arm selection in greedy

Select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

- Not much exploration!
  Explores once and then always makes the greedy choice.

Why Does Greedy Fail?

Arm selection in greedy

Select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

- Not much exploration!
  Explores once and then always makes the greedy choice.
- It can get stuck with a sub-optimal arm.

Why Does Greedy Fail?

Arm selection in greedy

Select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

- Not much exploration!
  Explores once and then always makes the greedy choice.
- It can get stuck with a sub-optimal arm.
- When?
  - the initial $\hat{\mu}$ of a sub-optimal arm is high, or
  - the initial $\hat{\mu}$ of the optimal arm is low.

Adding Exploration to Greedy

ϵ-Greedy : With probability $1 - \epsilon$,

choose the action with the highest empirical mean, and

with probability ϵ,

choose a random action.

Adding Exploration to Greedy

ϵ-Greedy : With probability $1-\epsilon$,

choose the action with the highest empirical mean, and

with probability ϵ,

choose a random action.

Algorithm ϵ-Greedy algorithm

```
for t = 1, ..., K do
    Choose each arm once.
end for
for t = K + 1, ... do
    Compute empirical means μ̂_1(t − 1), ..., μ̂_K(t − 1).
    With probability 1 − ε,
        select arm a(t) = arg max_a μ̂_a(t − 1).
    With probability ε,
        select a random arm.
end for
```

Adding Exploration to Greedy

ϵ-Greedy : With probability $1 - \epsilon$,
choose the action with the highest empirical mean, and
with probability ϵ,
choose a random action.

Algorithm ϵ-Greedy algorithm

for $t = 1, \ldots, K$ **do**
    Choose each arm once.
end for
for $t = K + 1, \ldots$ **do**
    Compute empirical means $\hat{\mu}_1(t-1), \ldots, \hat{\mu}_K(t-1)$.
    With probability $1 - \epsilon$,
        select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.
    With probability ϵ,
        select a random arm.
end for

ϵ-Greedy algorithm has linear regret!

Why Does ϵ-Greedy Fail?

Arm selection in ϵ-Greedy

With probability $1 - \epsilon$,

select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

With probability ϵ,

select a random arm.

Why Does ϵ-Greedy Fail?

Arm selection in ϵ-Greedy

With probability $1 - \epsilon$,

select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

With probability ϵ,

select a random arm.

- It explores forever.
- Constant ϵ ensures expected regret of at least

$$\sum_{a=1}^{K} \frac{\epsilon}{K} \Delta_a$$

at each time step.

Why Does ϵ-Greedy Fail?

Arm selection in ϵ-Greedy

With probability $1 - \epsilon$,

select arm $a(t) = \arg\max_a \hat{\mu}_a(t-1)$.

With probability ϵ,

select a random arm.

- It explores forever.
- Constant ϵ ensures expected regret of at least

$$\sum_{a=1}^{K} \frac{\epsilon}{K} \Delta_a$$

at each time step.

- Leading to expected cumulative regret of at least $\left(\frac{\epsilon}{K} \sum_{a=1}^{K} \Delta_a \right) T$.

Decaying ϵ-Greedy

- At time step t, explore with ϵ_t. A decay schedule for $\epsilon_1, \epsilon_2, \ldots$.

Decaying ϵ-Greedy

- At time step t, explore with ϵ_t. A decay schedule for $\epsilon_1, \epsilon_2, \ldots$.
- A schedule that has logarithmic regret: ☺

$$c > 0$$

$$d = \min_{a, \Delta_a > 0} \Delta_a$$

$$\epsilon_t = \min\left\{1, \frac{cK}{d^2 t}\right\}$$

Decaying ϵ-Greedy

- At time step t, explore with ϵ_t. A decay schedule for $\epsilon_1, \epsilon_2, \ldots$.
- A schedule that has logarithmic regret: ☺

$$c > 0$$
$$d = \min_{a, \Delta_a > 0} \Delta_a$$
$$\epsilon_t = \min\left\{1, \frac{cK}{d^2 t}\right\}$$

- Requires advance knowledge of gaps Δ ☹

Decaying ϵ-Greedy

- At time step t, explore with ϵ_t. A decay schedule for $\epsilon_1, \epsilon_2, \ldots$.
- A schedule that has logarithmic regret:

$$
\begin{aligned}
c &> 0 \\
d &= \min_{a, \Delta_a > 0} \Delta_a \\
\epsilon_t &= \min \left\{ 1, \frac{cK}{d^2 t} \right\}
\end{aligned}
$$

- Requires advance knowledge of gaps Δ
- Can we achieve sub-linear regret without such knowledge?

Break

We start again after a break.

Before the break

- Goal : Find algorithms with sub-linear regret.
- Greedy : Linear regret ☹
- ϵ-greedy : Linear regret ☹
- Decaying ϵ-greedy : Logarithmic regret, but requires advance knowledge of gaps Δ ☹
- Can we achieve sub-linear regret without such knowledge?

Optimism Principle

Optimism Principle informally

"You should act as if you are in the **best plausible** world."

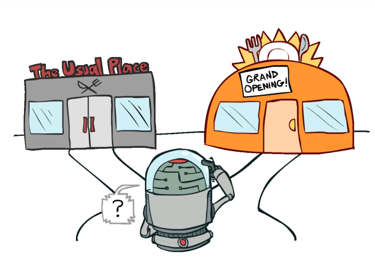


Image source: *UC Berkeley AI course, lecture 11*

Shall we try the new place?

Optimist : Yes!!! Pessimist : No!!!

Optimism Principle informally

"You should act as if you are in the **best plausible** world."

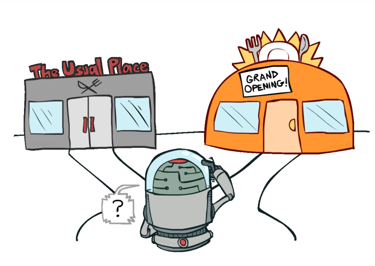


Image source: *UC Berkeley AI course, lecture 11*

Shall we try the new place?

Optimist : Yes!!! Pessimist : No!!!

Optimism guarantees either optimality or exploration.

Optimism Principle in Arm Selection

- Optimistic estimate of an arm = ‘Largest value it could plausibly be’.

Optimism Principle in Arm Selection

- Optimistic estimate of an arm = 'Largest value it could plausibly be'.
- 'Plausible'. The true mean cannot be *much larger* than the empirical mean.

Optimism Principle in Arm Selection

- Optimistic estimate of an arm = 'Largest value it could plausibly be'.
- 'Plausible'. The true mean cannot be *much larger* than the empirical mean.
- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) + \text{optimism term}$

Optimistic arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \text{optimism term}\right]$.

Optimism Principle in Arm Selection

- Optimistic estimate of an arm = 'Largest value it could plausibly be'.
- 'Plausible'. The true mean cannot be *much larger* than the empirical mean.
- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) +$ optimism term

  Similar to greedy, just with an addition of optimism term

Greedy arm selection

Select arm $a(t) = \arg\max_a [\hat{\mu}_a(t-1)]$.

Optimism Principle in Arm Selection

- Optimistic estimate of an arm = 'Largest value it could plausibly be'.
- 'Plausible'. The true mean cannot be *much larger* than the empirical mean.
- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) + \text{optimism term}$

Optimistic arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \text{optimism term}\right]$.

A Crash Course in Concentration of Measure

Concentration of Random Variables

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of of independent and identically distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 < \infty$.

$$\text{Empirical mean } \hat{\mu}_n = \frac{1}{n} \sum_{t=1}^{n} Z_t.$$

How close is $\hat{\mu}_n$ to μ?

Concentration of Random Variables

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of of independent and identically distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 < \infty$.

$$\text{Empirical mean } \hat{\mu}_n = \frac{1}{n} \sum_{t=1}^{n} Z_t.$$

How close is $\hat{\mu}_n$ to μ?

We could use law of large numbers

$$\lim_{n \to \infty} \hat{\mu}_n = \mu$$

Concentration of Random Variables

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of of independent and identically distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 < \infty$.

$$\text{Empirical mean } \hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} Z_t.$$

How close is $\hat{\mu}_n$ to μ?

We could use law of large numbers

$$\lim_{n\to\infty} \hat{\mu}_n = \mu$$

Law of large numbers requires $n \to \infty$. ☹

Preliminaries

Markov's inequality

If Z is a non-negative random variable and $c > 0$,

$$\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}.$$

Preliminaries

Markov's inequality

If Z is a non-negative random variable and $c > 0$,

$$\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}.$$

Subgaussian

Z is σ^2-subgaussian i.e. for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

Preliminaries

Markov's inequality

If Z is a non-negative random variable and $c > 0$,

$$\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}.$$

Subgaussian

Z is σ^2-subgaussian i.e. for all $\lambda \in \mathbb{R}$

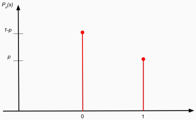
$$\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$$

Which distributions are σ-subgaussian? Gaussian, Bernoulli ....

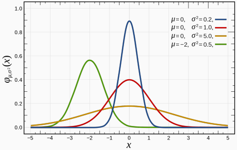
Recall : Distributional assumptions

Distributions $X_1, \ldots, X_K$ are unknown, we may make some assumptions:

- Bernoulli with unknown mean $\mu_a \in [0, 1]$.



- Gaussian with unit variance unknown mean $\mu_a \in \mathbb{R}$.



- Sub-Gaussian with unit variance.

Preliminaries

Markov's inequality

If Z is a non-negative random variable and $c > 0$,

$$\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$$

Z is sub-Gaussian with $\underline{\sigma^2 = 1}$.

Subgaussian

Z is σ^2-subgaussian i.e. for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$$

Which distributions are σ-subgaussian? Gaussian, Bernoulli . . . .

Preliminaries

Markov's inequality

If Z is a non-negative random variable and $c > 0$,

$$\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$$

Subgaussian

Z is 1-subgaussian i.e. for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$$

Which distributions are σ-subgaussian? Gaussian, Bernoulli . . . .

Concentration of sub-Gaussian random variables

Chernoff-Hoeffding bound

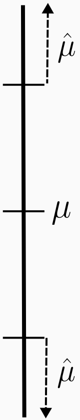
Let $Z_1, \dots Z_n$ are independent sub-Gaussian random variables with mean μ and variance 1 and,

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} Z_t,$$

then for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$$

$$\mathbb{P}\left(\hat{\mu} \leq \mu - \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$$



Recall : optimism principle in arm selection

- Optimistic estimate of an arm = Largest value it could plausibly be.
- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) +$ optimism term

Optimistic arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \text{optimism term}\right]$.

Optimism term of the form $\sqrt{\frac{2\log(1/\delta)}{n}}$?

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right)$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n} (Z_t - \mu) \geq \epsilon n\right)$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp\left(\lambda \epsilon n\right)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

(1) $\mathbb{P}(Z \geq c) \leq \dfrac{\mathbb{E}[Z]}{c}$

(2) $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp(\lambda \epsilon n)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp(-\lambda \epsilon n) \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality (1)}$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp\left(\lambda \epsilon n\right)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp\left(-\lambda \epsilon n\right) \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality } ①$$

$$= \exp\left(-\lambda \epsilon n\right) \cdot \mathbb{E}\left[\prod_{t=1}^{n} \exp\left(\lambda\left(Z_t - \mu\right)\right)\right]$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

① $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$

② $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda\sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp\left(\lambda\epsilon n\right)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp\left(-\lambda\epsilon n\right) \cdot \mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality ①}$$

$$= \exp\left(-\lambda\epsilon n\right) \cdot \mathbb{E}\left[\prod_{t=1}^{n}\exp\left(\lambda\left(Z_t - \mu\right)\right)\right] \leq \exp\left(-\lambda\epsilon n\right) \cdot \prod_{t=1}^{n}\exp\left(\lambda^2/2\right)$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1 $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$

2 $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp(\lambda \epsilon n)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp(-\lambda\epsilon n) \cdot \mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality } ①$$

$$= \exp(-\lambda\epsilon n) \cdot \mathbb{E}\left[\prod_{t=1}^{n}\exp(\lambda(Z_t - \mu))\right] \leq \exp(-\lambda\epsilon n) \cdot \prod_{t=1}^{n}\exp\left(\lambda^2/2\right)$$

$$= \exp\left(-\lambda\epsilon n + \frac{\lambda^2 n}{2}\right)$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp\left(\lambda \epsilon n\right)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp\left(-\lambda \epsilon n\right) \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality } ①$$

$$= \exp\left(-\lambda \epsilon n\right) \cdot \mathbb{E}\left[\prod_{t=1}^{n}\exp\left(\lambda\left(Z_t - \mu\right)\right)\right] \leq \exp\left(-\lambda \epsilon n\right) \cdot \prod_{t=1}^{n}\exp\left(\lambda^2/2\right)$$

$$= \exp\left(-\lambda \epsilon n + \frac{\lambda^2 n}{2}\right) = \exp\left(-\frac{\epsilon^2 n}{2}\right) \qquad \text{for } \lambda = \epsilon$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp(\lambda \epsilon n)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp(-\lambda\epsilon n) \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality } (1)$$

$$= \exp(-\lambda\epsilon n) \cdot \mathbb{E}\left[\prod_{t=1}^{n} \exp(\lambda(Z_t - \mu))\right] \leq \exp(-\lambda\epsilon n) \cdot \prod_{t=1}^{n} \exp(\lambda^2/2)$$

$$= \exp\left(-\lambda\epsilon n + \frac{\lambda^2 n}{2}\right) = \exp\left(-\frac{\epsilon^2 n}{2}\right) \qquad \text{for } \lambda = \epsilon$$

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) \leq \exp\left(-\frac{\epsilon^2 n}{2}\right)$$

Proving Chernoff-Hoeffding bound

To prove: $\mathbb{P}\left(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta$

1. $\mathbb{P}(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}$
2. $\mathbb{E}[\exp(\lambda Z)] \leq \exp\left(\frac{\lambda^2}{2}\right)$

Proof:

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{n}(Z_t - \mu) \geq \epsilon n\right)$$

$$= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right) \geq \exp(\lambda \epsilon n)\right) \qquad \text{for some } \lambda \in \mathbb{R}$$

$$\leq \exp(-\lambda \epsilon n) \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n}(Z_t - \mu)\right)\right] \qquad \text{by Markov's inequality } (1)$$

$$= \exp(-\lambda \epsilon n) \cdot \mathbb{E}\left[\prod_{t=1}^{n} \exp(\lambda(Z_t - \mu))\right] \leq \exp(-\lambda \epsilon n) \cdot \prod_{t=1}^{n} \exp(\lambda^2/2)$$

$$= \exp\left(-\lambda \epsilon n + \frac{\lambda^2 n}{2}\right) = \exp\left(-\frac{\epsilon^2 n}{2}\right) \qquad \text{for } \lambda = \epsilon$$

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) \leq \exp\left(-\frac{\epsilon^2 n}{2}\right) \qquad \epsilon = \sqrt{\frac{2\log(1/\delta)}{n}} \rightarrow \delta = \exp\left(-\frac{\epsilon^2 n}{2}\right)$$

Upper Confidence Bound (UCB) algorithm

Upper Confidence Bound (UCB) : Choose Arms Optimistically

- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) +$ optimism term
- Optimism term of the form $\sqrt{\frac{2\log(1/\delta)}{n}}$?

Optimistic arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \text{optimism term}\right]$.

Upper Confidence Bound (UCB) : Choose Arms Optimistically

- Optimistic estimate of arm $a = \hat{\mu}_a(t-1) +$ optimism term
- UCB estimate of arm $a = \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

Upper Confidence Bound (UCB) : Choose arms optimally

Algorithm UCB algorithm Auer et al. [2002]

Parameters : Confidence level δ

for $t = 1, \ldots, K$ **do**
  Choose each arm once.
end for
for $t = K + 1, \ldots$ **do**
  Compute empirical means $\hat{\mu}_1(t-1), \ldots, \hat{\mu}_K(t-1)$.
  Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.
end for

Regret bound for UCB

Theorem

The expected cumulative regret of UCB after T time steps is

$$\textit{Regret} = \mathfrak{R}(T) \leq \sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} + 3\Delta_a.$$

Logarithmic regret 🙂

Proving the Regret Bound for UCB : Roadmap

- Decomposition of regret over the arms.

Proving the Regret Bound for UCB : Roadmap

- Decomposition of regret over the arms.
- On a 'good' event, prove that sub-optimal arms are not played too often.

Proving the Regret Bound for UCB : Roadmap

- Decomposition of regret over the arms.
- On a ‘good’ event, prove that sub-optimal arms are not played too often.
- Prove that the ‘good’ event occurs with a high probability.

Proving the Regret Bound for UCB : Roadmap

- Decomposition of regret over the arms. $\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$

  where $\Delta_a \coloneqq \mu_* - \mu_a$ and $N_a(T) \coloneqq \sum_{t=1}^{T} \mathbb{I}(a(t) = a)$
- On a 'good' event, prove that sub-optimal arms are not played too often.
- Prove that the 'good' event occurs with a high probability.

Proving the Regret Bound for UCB : I

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

'Good event': When UCB performs well.

Fix a sub-optimal arm a. Assume for all t,

Proving the Regret Bound for UCB : I

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

'Good event': When UCB performs well.

Fix a sub-optimal arm a. Assume for all t,

Empirical estimate of sub-optimal arm a is not too big.

Proving the Regret Bound for UCB : I

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

'Good event': When UCB performs well.

Fix a sub-optimal arm a. Assume for all t,

Empirical estimate of sub-optimal arm a is not too big.

$$\hat{\mu}_a(t-1) \leq \mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}.$$

Proving the Regret Bound for UCB : I

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

'Good event': When UCB performs well.

Fix a sub-optimal arm a. Assume for all t,

Empirical estimate of sub-optimal arm a is not too big.

$$\hat{\mu}_a(t-1) \leq \mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}.$$

Empirical estimate of optimal arm a_* is not too small.

Proving the Regret Bound for UCB : I

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

'Good event': When UCB performs well.

Fix a sub-optimal arm a. Assume for all t,

Empirical estimate of sub-optimal arm a is not too big.

$$\hat{\mu}_a(t-1) \leq \mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}.$$

Empirical estimate of optimal arm a_* is not too small.

$$\hat{\mu}_{a_*}(t-1) \geq \mu_* - \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}}.$$

Proving the Regret Bound for UCB : II

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

1. $\mu_a \; + \; \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1)$,
2. $\hat{\mu}_{a_*}(t-1) \; + \; \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \geq \mu_*$.

At time t, the algorithm selects a only if,

Proving the Regret Bound for UCB : II

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

1. $\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1)$,
2. $\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \geq \mu_*$.

At time t, the algorithm selects a only if,

$$\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \qquad \text{using } 1$$

Proving the Regret Bound for UCB : II

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

1 $\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1)$,

2 $\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \geq \mu_*$.

At time t, the algorithm selects a only if,

$$
\begin{aligned}
\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} &\geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \qquad \text{using } 1 \\
&\geq \hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}}
\end{aligned}
$$

Proving the Regret Bound for UCB : II

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

1 $\mu_a \ + \ \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1)$,

2 $\hat{\mu}_{a_*}(t-1) \ + \ \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \geq \mu_*$.

At time t, the algorithm selects a only if,

$$
\begin{aligned}
\mu_a \ + \ 2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} &\geq \hat{\mu}_a(t-1) \ + \ \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \qquad \text{using } 1 \\
&\geq \hat{\mu}_{a_*}(t-1) \ + \ \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \\
&\geq \mu_*
\end{aligned}
$$

Proving the Regret Bound for UCB : II

UCB arm selection

Select arm $a(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

1 $\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \hat{\mu}_a(t-1)$,

2 $\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \geq \mu_*$.

At time t, the algorithm selects a only if,

$$
\begin{aligned}
\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} &\geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} && \text{using } 1 \\
&\geq \hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \\
&\geq \mu_* = \mu_a + \Delta_a && \text{using } 2
\end{aligned}
$$

$$
\cancel{\mu_a} + 2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \cancel{\mu_a} + \Delta_a
$$

Proving the Regret Bound for UCB : III

If the good event occurs,
at time t, the algorithm selects a only if,

$$2\sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \geq \Delta_a$$

$$N_a(t-1) \leq \frac{8\log(1/\delta)}{\Delta_a^2}$$

So assuming the good event occurs,

$$N_a(T) \leq \frac{8\log(1/\delta)}{\Delta_a^2} + 1.$$

Probability (Good Event Does Not Occur)

The good event,

$$
\begin{aligned}
\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} &\geq \hat{\mu}_a(t-1) \\
\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} &\geq \mu_*
\end{aligned}
$$

Probability (Good Event Does Not Occur)

The good event does not occur,

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \leq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \leq \mu_*$$

Probability (Good Event Does Not Occur)

The good event does not occur at time step t,

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \leq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \leq \mu_*$$

Chernoff-Hoeffding bound shows that

$$\mathbb{P}\left(\hat{\mu}_a(t-1) \geq \mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right) \leq \delta$$

$$\mathbb{P}\left(\hat{\mu}_{a_*}(t-1) \leq \mu_* - \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}}\right) \leq \delta$$

Probability (Good Event Does Not Occur)

The good event does not occur at some step t, $1 \leq t \leq T$,

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} \leq \hat{\mu}_a(t-1)$$

$$\hat{\mu}_{a_*}(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(t-1)}} \leq \mu_*$$

Chernoff-Hoeffding bound combined with union bound $\mathbb{P}(\cup_i E_i) \leq \sum_i \mathbb{P}(E_i)$,

$$\mathbb{P}\left(\exists \tau \leq T : \hat{\mu}_a(\tau-1) \geq \mu + \sqrt{\frac{2\log(1/\delta)}{N_a(\tau-1)}}\right) \leq \delta T$$

$$\mathbb{P}\left(\exists \tau \leq T : \hat{\mu}_{a_*}(\tau-1) \leq \mu_* - \sqrt{\frac{2\log(1/\delta)}{N_{a_*}(\tau-1)}}\right) \leq \delta T$$

Proving the Regret Bound for UCB : IV

1. $N_a(T) \leq \frac{8\log(1/\delta)}{\Delta_a^2} + 1$ when the good event occurs.
2. Probability (good event does not occur) $\leq 2\delta T$.

Using the decomposition of regret $\mathfrak{R}(T)$ over the arms,

$$\mathfrak{R}(T) = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$$

Proving the Regret Bound for UCB : IV

1. $N_a(T) \leq \frac{8\log(1/\delta)}{\Delta_a^2} + 1$ when the good event occurs.
2. Probability (good event does not occur) $\leq 2\delta T$.

Using the decomposition of regret $\mathfrak{R}(T)$ over the arms,

$$\begin{aligned}
\mathfrak{R}(T) &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\
&\leq \sum_{a:\Delta_a>0} \Delta_a \left[\frac{8\log(1/\delta)}{\Delta_a^2} + 1 + 2\delta T \cdot T \right]
\end{aligned}$$

Proving the Regret Bound for UCB : IV

1. $N_a(T) \leq \frac{8\log(1/\delta)}{\Delta_a^2} + 1$ when the good event occurs.
2. Probability (good event does not occur) $\leq 2\delta T$.

Using the decomposition of regret $\mathfrak{R}(T)$ over the arms,

$$\begin{aligned}
\mathfrak{R}(T) &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\
&\leq \sum_{a:\Delta_a>0} \Delta_a \left[\frac{8\log(1/\delta)}{\Delta_a^2} + 1 + 2\delta T \cdot T\right] \\
&\leq \sum_{a:\Delta_a>0} \Delta_a \left[\frac{8\log(T^2)}{\Delta_a^2} + 1 + 2\frac{1}{T^2}T^2\right] \qquad \text{choosing } \delta = 1/T^2,
\end{aligned}$$

Proving the Regret Bound for UCB : IV

1. $N_a(T) \le \frac{8\log(1/\delta)}{\Delta_a^2} + 1$ when the good event occurs.
2. Probability (good event does not occur) $\le 2\delta T$.

Using the decomposition of regret $\mathfrak{R}(T)$ over the arms,

$$
\begin{aligned}
\mathfrak{R}(T) &= \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)] \\
&\le \sum_{a:\Delta_a>0} \Delta_a \left[\frac{8\log(1/\delta)}{\Delta_a^2} + 1 + 2\delta T \cdot T\right] \\
&\le \sum_{a:\Delta_a>0} \Delta_a \left[\frac{8\log(T^2)}{\Delta_a^2} + 1 + 2\frac{1}{T^2} T^2\right] \qquad \text{choosing } \delta = 1/T^2, \\
&= \sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} + 3\Delta_a.
\end{aligned}
$$

Regret Bound for UCB

Theorem

The expected cumulative regret of UCB after T time steps is

$$Regret = \mathfrak{R}(T) \leq \sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} + 3\Delta_a.$$

Regret Bound for UCB

Theorem

The expected cumulative regret of UCB after T time steps is

$$\textit{Regret} = \mathfrak{R}(T) \leq \sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} + 3\Delta_a.$$

Distribution-dependent regret bound.

Distribution-free Regret Bound for UCB

$$\mathfrak{R}(T) = \sum_{a:\Delta_a > 0} \Delta_a \, \mathbb{E}[N_a(T)]$$

Distribution-free Regret Bound for UCB

$$
\begin{aligned}
\mathfrak{R}(T) &= \sum_{a:\Delta_a>0} \Delta_a \, \mathbb{E}[N_a(T)] \\
&= \sum_{a:\Delta_a>0,\Delta_a\leq\Delta} \Delta_a \, \mathbb{E}[N_a(T)] + \sum_{a:\Delta_a>\Delta} \Delta_a \, \mathbb{E}[N_a(T)]
\end{aligned}
$$

Distribution-free Regret Bound for UCB

$$\begin{aligned}
\mathfrak{R}(T) &= \sum_{a:\Delta_a>0} \Delta_a \, \mathbb{E}[N_a(T)] \\
&= \sum_{a:\Delta_a>0,\Delta_a\leq\Delta} \Delta_a \, \mathbb{E}[N_a(T)] + \sum_{a:\Delta_a>\Delta} \Delta_a \, \mathbb{E}[N_a(T)] \\
&\leq \Delta T + \sum_{a:\Delta_a>\Delta} \frac{16\log(T)}{\Delta_a} + 3\Delta_a
\end{aligned}$$

Distribution-free Regret Bound for UCB

$$
\begin{aligned}
\mathfrak{R}(T) &= \sum_{a:\Delta_a>0} \Delta_a \, \mathbb{E}[N_a(T)] \\
&= \sum_{a:\Delta_a>0,\Delta_a\leq\Delta} \Delta_a \, \mathbb{E}[N_a(T)] + \sum_{a:\Delta_a>\Delta} \Delta_a \, \mathbb{E}[N_a(T)] \\
&\leq \Delta T + \sum_{a:\Delta_a>\Delta} \frac{16\log(T)}{\Delta_a} + 3\Delta_a \\
&\leq O(\sqrt{KT\log(T)}) \qquad \text{using } \Delta = \sqrt{K\log T/T}.
\end{aligned}
$$

A primer on *big-oh* notation $O(.)$

Summary

- Stationary stochastic bandits.

Summary

- Stationary stochastic bandits.
- Why greedy and ϵ-greedy does not work?

Summary

- Stationary stochastic bandits.
- Why greedy and ϵ-greedy does not work?
- A short introduction to concentration of measure.

Summary

- Stationary stochastic bandits.
- Why greedy and ϵ-greedy does not work?
- A short introduction to concentration of measure.
- UCB algorithm and its regret bound.

Next lecture

- Bayesian way of looking at bandits.

Next lecture

- Bayesian way of looking at bandits.
- Leading to another algorithm and its regret bound.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.