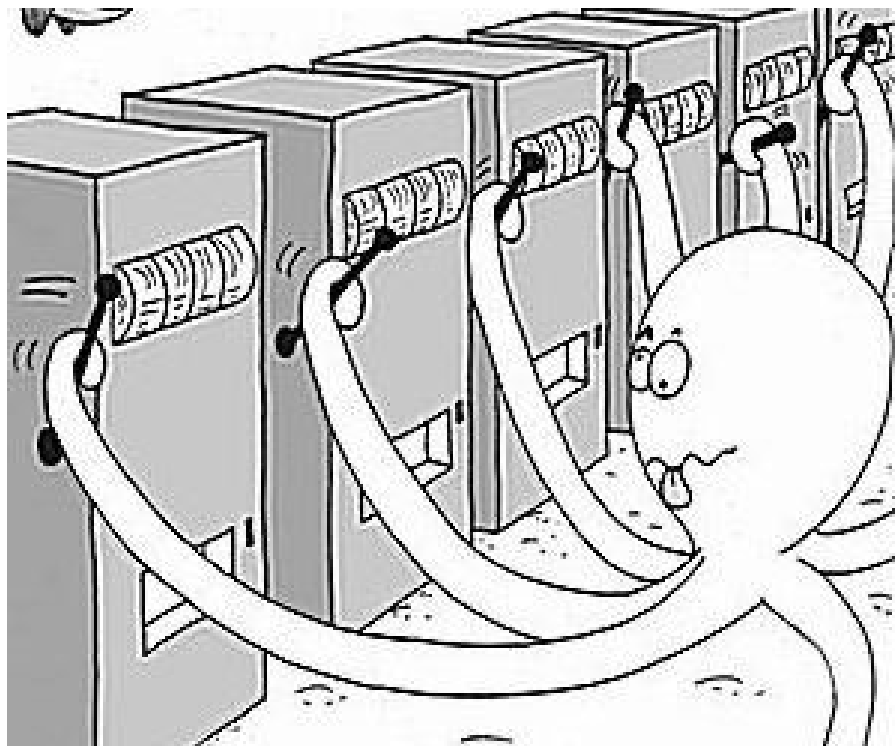


Multi-armed Bandits

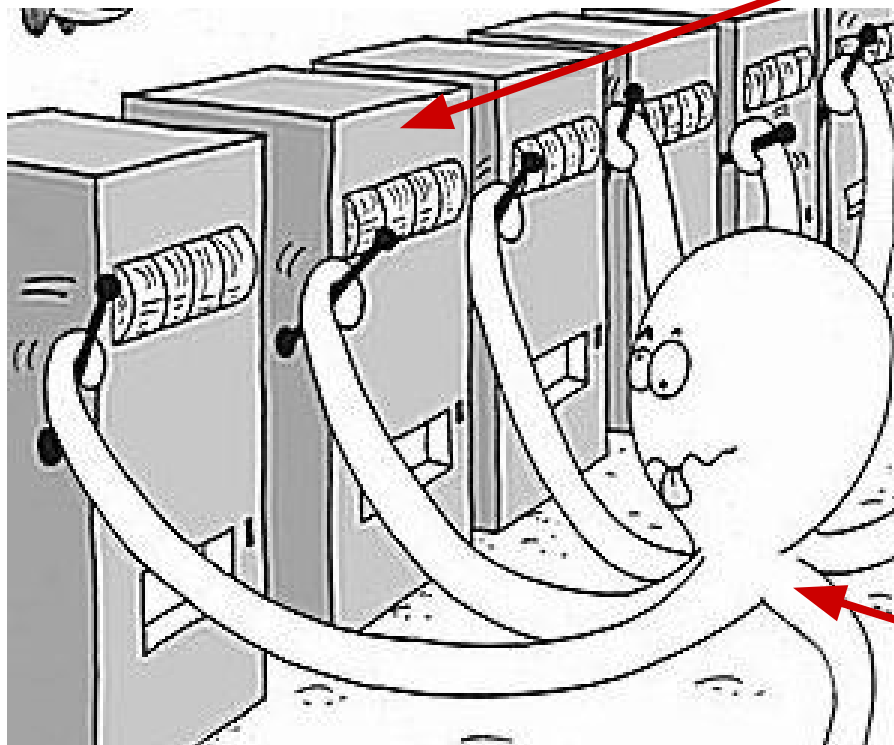
Connor Lee, Hoang Le, Ritvik Mishra

Multi-Armed Bandits Problem



Multi-Armed Bandits Problem

n slot machines ("bandits")

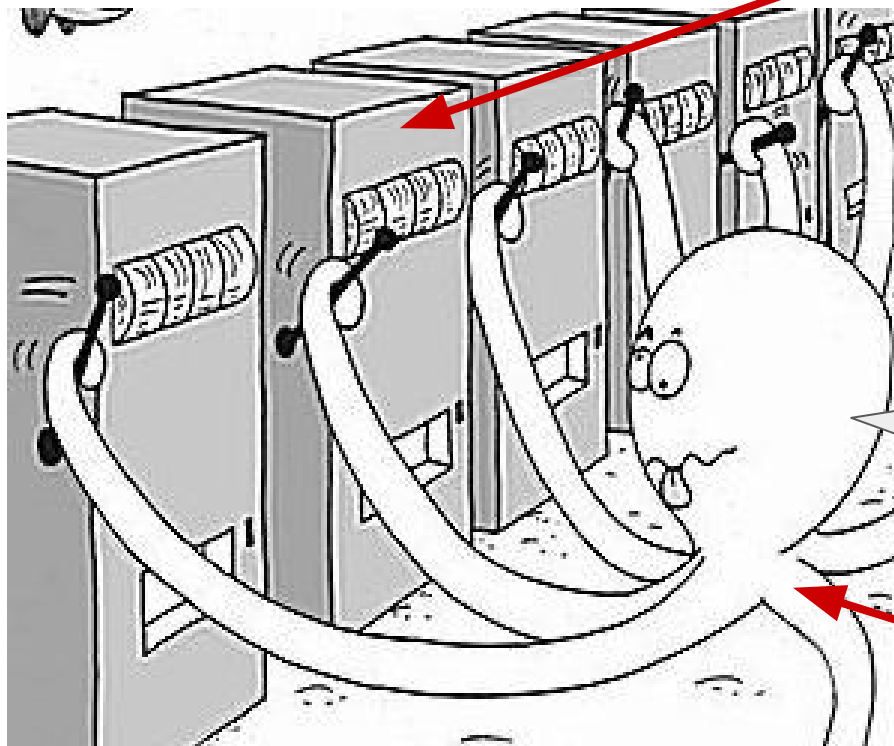


gambler

Multi-Armed Bandits Problem

n slot machines (“bandits”)

12:00 PM



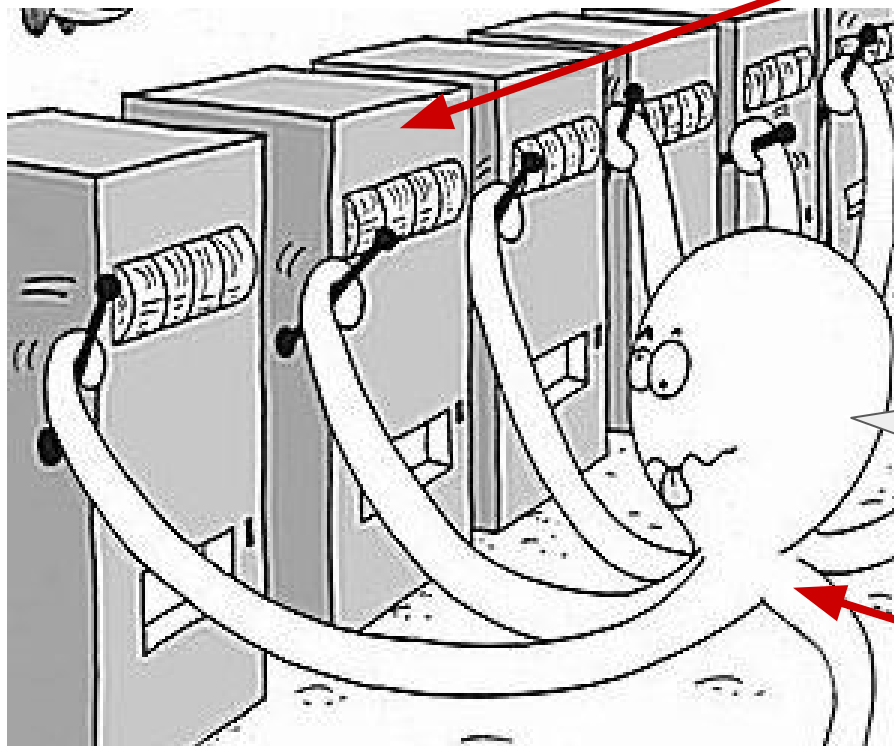
Wow! I earned nothing from the machine 1!

gambler

Multi-Armed Bandits Problem

n slot machines (“bandits”)

12:01 PM



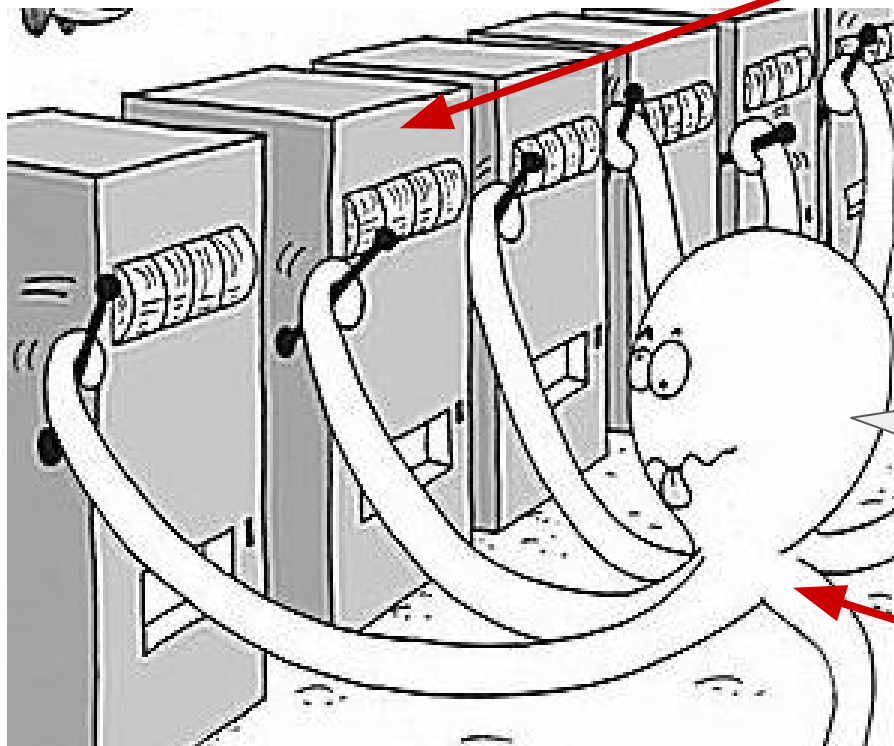
Wow! I earned \$10 from the machine 2.

gambler

Multi-Armed Bandits Problem

n slot machines (“bandits”)

12:02 PM



Wow! I earned \$2 from the machine 5.

gambler

Multi-Armed Bandits Problem

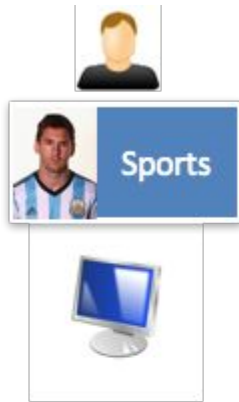
Gambler has a row of n slot machines.

At each time step $t = 0 \dots T$, choose a slot machine to play.

Experiences loss *only* from the attempted slot machine.


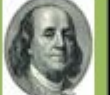



Does not know what would have happened had another slot machine been chosen.

Example: Interactive News Recommender (5 Classes, No Features)



Average Likes

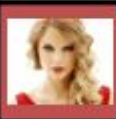
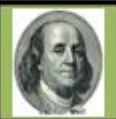



Shown

				
--	--	--	--	--
0	0	0	1	0



Example: Interactive News Recommender (5 Classes, No Features)



					
Average Likes	--	--	--	0	--
# Shown	0	0	0	1	0





Example: Interactive News Recommender (5 Classes, No Features)



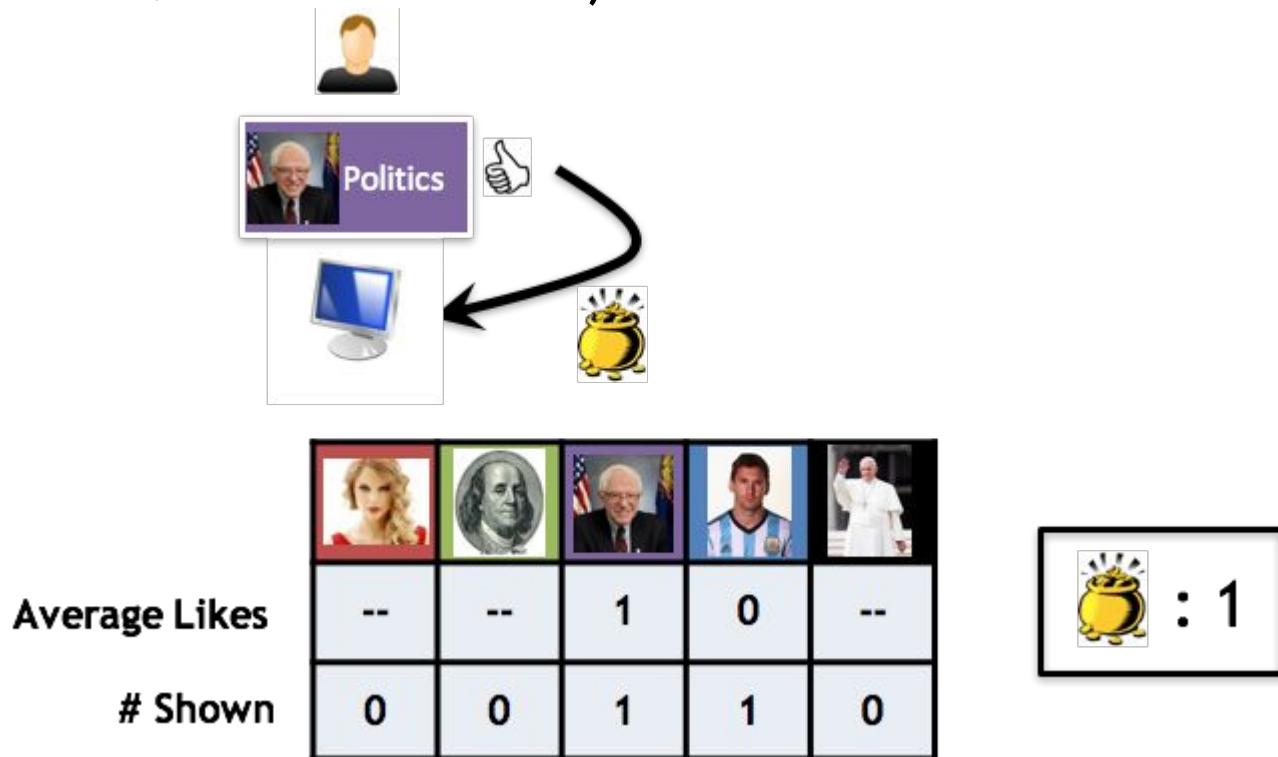
Average Likes

Shown

					
Average Likes	--	--	--	0	--
# Shown	0	0	1	1	0



Example: Interactive News Recommender (5 Classes, No Features)

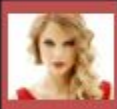


Example: Interactive News Recommender (5 Classes, No Features)



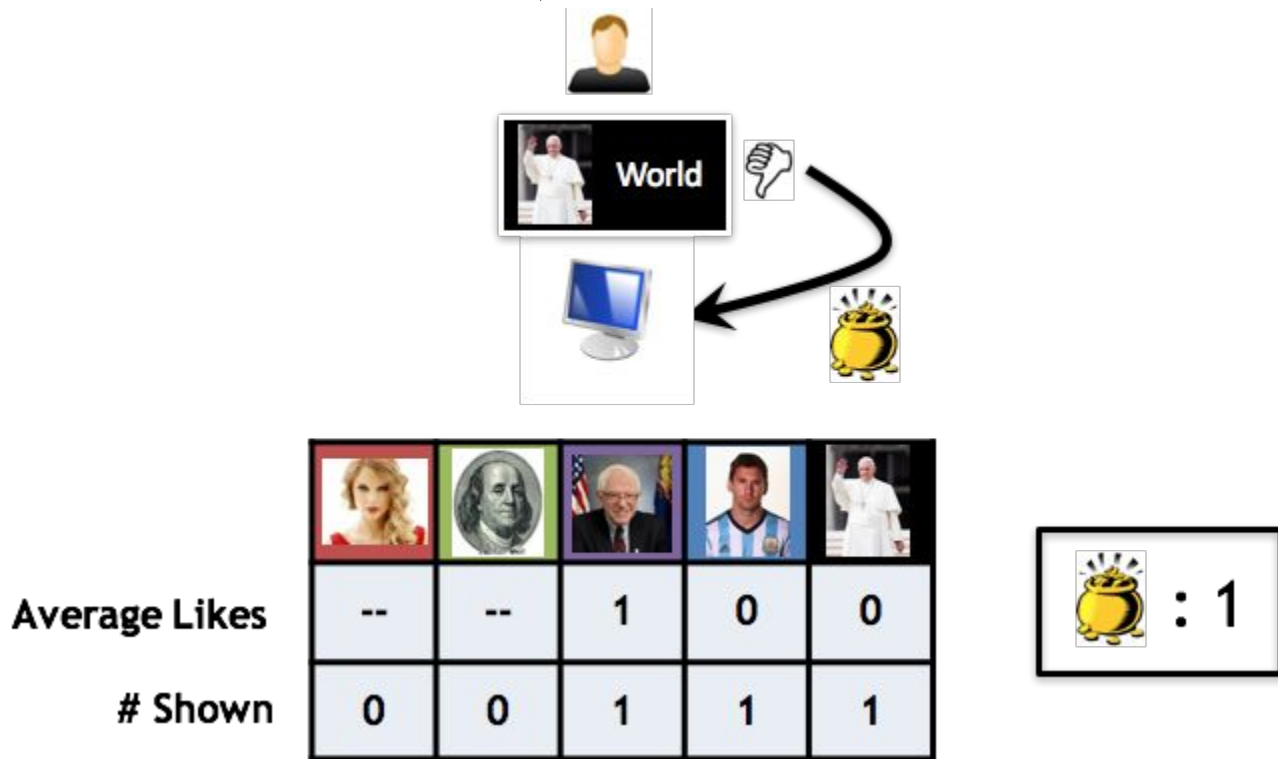
Average Likes

Shown

				
--	--	1	0	--
0	0	1	1	1



Example: Interactive News Recommender (5 Classes, No Features)


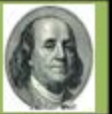





Example: Interactive News Recommender (5 Classes, No Features)



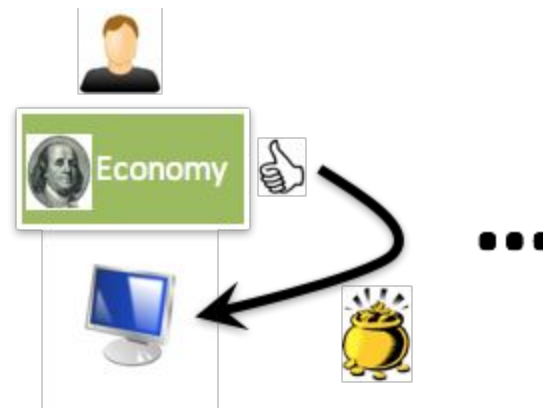
Average Likes

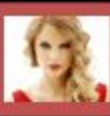
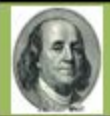



Shown

				
--	--	1	0	0
0	1	1	1	1



Example: Interactive News Recommender (5 Classes, No Features)



					
Average Likes	--	1	1	0	0
# Shown	0	1	1	1	1



What Should Algorithm Recommend

Exploit:



Explore:




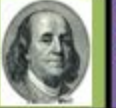



Best:



How to Optimally Balance Explore/Exploit Tradeoff?
Characterized by the Multi-Armed Bandit Problem

Average Likes

Shown

					
Average Likes	--	0.44	0.4	0.33	0.2
# Shown	0	25	10	15	20



: 24

Efficiency Measure = Regret

$$\text{💰}(OPT) = \text{💰}(\text{🇺🇸}) + \text{💰}(\text{🇺🇸}) + \text{💰}(\text{🇺🇸}) \dots$$

$$\text{💰}(ALG) = \text{💰}(\text{🇦🇷}) + \text{💰}(\text{🇺🇸}) + \text{💰}(\text{🇵🇪}) \dots$$

Time Horizon

Regret: $R(T) = \text{💰}(OPT) - \text{💰}(ALG)$

- Opportunity cost of not knowing preferences
- “no-regret” if $R(T)/T \rightarrow 0$
 - Efficiency measured by convergence rate

Formal Definition

K actions/classes

Each action a has an average reward: μ_i , where $1 \leq i \leq K$

For $t = 1 \dots T$:

Choose action a_t

Receive reward $X_{i,n}$

Goal: Minimize Expected Regret

$$\mu^* n - \sum_{j=1}^K \mu_j \mathbb{E}[T_j(n)] \quad \text{where } \mu^* \triangleq \max_{1 \leq i \leq K} \mu_i$$

Formal Definition

Each action a has an average reward: μ_i , where $1 \leq i \leq K$

True averages for all arms are not known.

Only the reward associated with chosen arm is observed.

Weighted Majority	Multiplicative Weights	Multi-armed Bandits
Full Information	Full Information	Partial Information

Two Main Branches of Bandit Problems

Stochastic bandits

The reward $X_{i,n}$ is sampled from an unknown product distribution $\nu_1 \otimes \dots \otimes \nu_k$ on $[0, 1]^k$ and $X_{i,n} \sim \nu_i$.

Adversarial bandits

$X_{i,n}$ is chosen by an adversary. Adversary knows the strategy we are employing and the history, but **does not know** the action we take at time n prior to forming $X_{i,n}$.

Comparison with other Online Learning Settings

- Can only know the loss of one arm at each timestep. *“Partial Information”*
- There exist algorithms where knowledge of horizon T is not required - *“Anytime Algorithm”*
- Basic stochastic multi-armed bandits have **no features**.

We will primarily consider **stochastic bandits**.

Key theme: *exploration vs. exploitation*

Extensions of Multi-armed Bandit Problems

- We can consider (featurized) observations of the world when making a decision. This is known as the *contextual bandit* problem
- UCB1 is the building block for tree search algorithms (e.g. UCT, Monte Carlo Tree Search) used to learn to play games (e.g. Go)
- Considering the effect of sequence of decisions (i.e. allowing decisions to effect the world) is the general problem of *reinforcement learning*

Exploitation vs. Exploration

For **stochastic bandits**, there is some empirically found mean reward for each of the arms.

We want to find arm with highest true mean reward.

Exploitation vs. Exploration

Exploitation

Choose the arm with the highest empirical mean reward.

Exploration

Test other arms to potentially higher actual mean reward.

How do we know when to exploit and when to explore?

Thought Experiment: Exploit-only

1. Find the average reward of all arms by picking each a few times.
2. Exploit the arm with highest mean reward.

At timestep i , what if there exists another arm with much greater reward that was not sufficiently explored?

To be confident that this other arm does not exist, we would need to test all of the arms *many* times initially. This is exploring, in fact it is over-exploring!

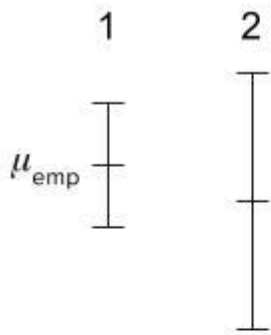
Thought Experiment: Exploit-only

Imagine you have two arms:

- One with constant reward of 10
- And the other with 99% probability of 0 reward, 1% probability of 10000 reward

If you initially pull each level k times, there is a $99\%^k$ chance of the exploit-only solution resulting in a very suboptimal strategy!

Thought Experiment: Incorporate Exploration



You have these three bandits. The bars represent the empirical mean, \pm the uncertainty. Assume the true mean is within these ranges.

Which bandit should you choose?

If you do “pure exploitation”: you choose bandit 1, and run the risk of bandit 2 having a higher true mean.

If you do “pure exploration”: you choose bandit 3 since it has the largest uncertainty, even though it clearly cannot be better than bandit 1.

Thought Experiment: Incorporate Exploration



Clearly, need an intermediate strategy that allows you to explore bandit 2 since it still has a large uncertainty and there is a chance that it is better than bandit 1.

Perhaps you choose the bandit with the largest mean + uncertainty?

There is an upper bound to how much worse bandit 2 can be relative to bandit 1 based on the uncertainties.

UCB1 Algorithm

UCB1 Algorithm

Initialization: Play each action once.

Loop at each round n : Play action j that maximizes

$$\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$$

Average reward observed from j

Index of round, so total number of attempts thus far

Number of times j has been played so far

UCB1 Confidence Interval

Estimate of Expected
Reward from data


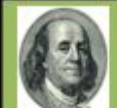



$$\bar{x}_j \pm \sqrt{\frac{2 \ln n}{n_j}}$$

Number of rounds so far
(70 in example below)

#times arm j was chosen

average likes

#times chosen

					
average likes	--	0.44	0.4	0.33	0.2
#times chosen	0	25	10	15	20

Confidence Interval

- Maintain confidence interval for each action
 - In the UCB1 case: derived using Chernoff-Hoeffding bounds




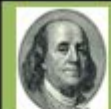



= [-0.1, 1.0]



= [-0.5, 1.3]

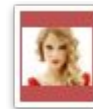


= [-0.4, 1.1]

					
average likes	--	0.44	0.4	0.33	0.2
#times chosen	0	25	10	15	20



= [-0.5, 0.8]



Undefined


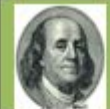



Balancing Exploration / Exploitation

➤ Optimism in the Face of Uncertainty

- At any time n , from past observations and probabilistic derivations, we have an upper confidence bound on the expected rewards
- Simple implementation:
 - Play the arm having the current largest UCB

$$\underset{j}{\operatorname{argmax}} \bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$$

Exploitation Term Exploration Term

					
average likes	--	0.44	0.4	0.33	0.2
#times chosen	0	25	10	15	20

Thought Experiment

➤ Could we stay a long time taking a wrong action?

- No, because:

- The more we draw a wrong arm j the closer UCB gets to the expected reward μ_j

$$\mu_j < \mu^* \leq \text{UCB on } \mu^*$$


- Number of times sub-optimal action is taken $O\left(\frac{\ln n}{(\mu^* - \mu_j)^2}\right)$

Thought Experiment


- What if a good action never gets taken?
 - An arm never gets selected if:

$$\mu_j + \sqrt{(2 \ln n) / n_j} \leq \mu^*$$

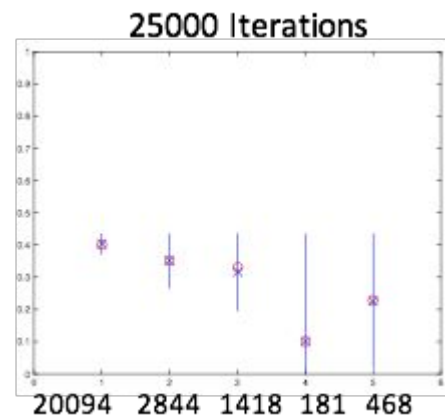
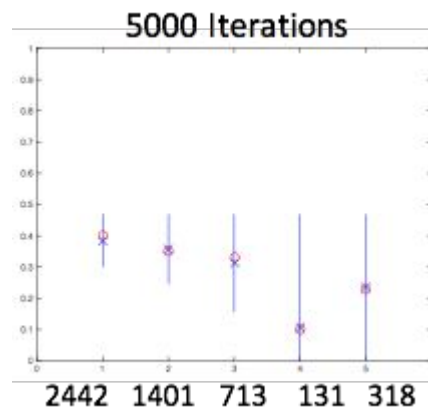
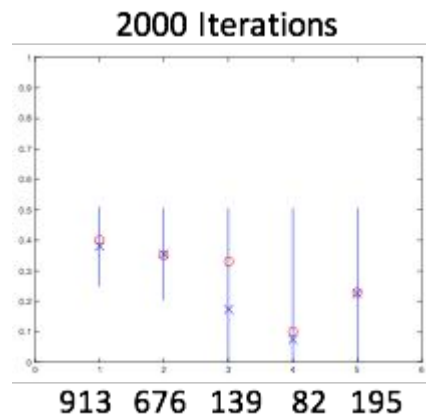
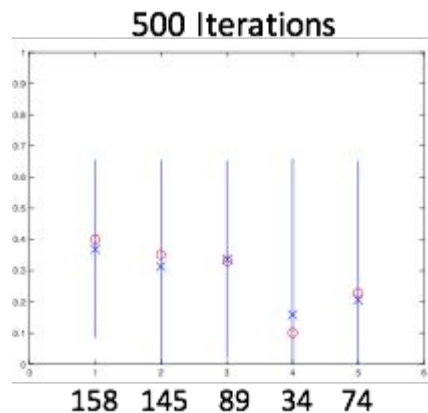
Bound grows
slowly with time



Shrinks quickly
with #trials



Simulation



High-Level Intuition of Analysis

At current round, chosen arm $j = \operatorname{argmax}_i \bar{x}_i + \sqrt{\frac{2 \ln n}{n_i}}$

Can show with high probability:

$$\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}} \geq \bar{x}^* + \sqrt{\frac{2 \ln n}{n^*}} \geq \mu^*$$

Upper Confidence Bound of Best Arm

Value of Best Arm

$$\mu_j \geq \bar{x}_j - \sqrt{\frac{2 \ln n}{n_j}}$$

The true value is greater than the lower confidence bound.

$$\implies \mu^* - \mu_j \leq 2\sqrt{\frac{2 \ln n}{n_j}}$$

Bound on regret at time n less than twice the size of confidence interval

Balancing Exploitation vs. Exploration in UCB1

UCB1 chooses the bandit with the maximum empirical mean + confidence interval = upper confidence bound (UCB).

Confidence interval is defined so that real mean of chosen bandit is **at least 2 confidence intervals less than the optimal payoff** (with high probability).

Exploration: less-explored bandits have large confidence intervals, so they will be chosen until their UCB is too low to be chosen again.

Exploitation: confidence intervals shrink and the high empirical mean bandits are chosen.

Regret Bound of UCB1 - Formal Statement

Theorem: For $K > 1$, if UCB1 is run on K machines having arbitrary reward distributions P_1, \dots, P_K with support in $[0, 1]$, then its expected regret after any number n of plays is at most

$$\left[8 \sum_{i: \mu_i < \mu^*} \left(\frac{\ln n}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^K \Delta_j \right)$$

where $\Delta_i \triangleq \mu^* - \mu_i$ is positive constant (unknown) gap

Regret Bound of UCB1

Takeaway: logarithmic regret, anytime algorithm

In comparison with previous online learning setting, with high probability, UCB1 accumulates regret at most:

$$R(T) = O\left(\frac{K}{\epsilon} \ln T\right)$$

#Actions

Time Horizon

Gap between best & 2nd best
 $\epsilon = \mu^* - \mu_2$

Proof of Theorem 1 - Notations

➤ $\bar{X}_{j,n}$ = empirical average of j after action j has been taken n times

$$\bar{x}_j \triangleq \bar{X}_{j,n} = \frac{1}{n} \sum_{t=1}^n X_{j,t}$$

➤ $T_j(n)$ = number of times arm j played in the first n plays

$$\sum_{j=1}^K T_j(n) = n$$

Proof of Theorem 1 - Notations

Action j selected at time n , shorthand notation for UCB:

$$\begin{aligned} U_j(n) &= \bar{x}_j + \sqrt{\frac{2 \ln n}{T_j(n)}} \\ &= \bar{X}_{j,T_j(n)} + c_{n,T_j(n)} \text{ with } c_{n,s} = \sqrt{(2 \ln n)/s} \end{aligned}$$

Random variable I_t denotes the arm played at time $t \quad \forall t$

Detailed Proof of Theorem 1

$$\sum_{j=1}^K T_j(n) = n$$

Rewrite:


$$\begin{aligned} \mathbf{Regret} &= \mu^* n - \sum_{j=1}^K \mu_j \mathbb{E}[T_j(n)] \quad \text{where } \mu^* \triangleq \max_{1 \leq i \leq K} \mu_i \\ &= \sum_{j: \mu_j < \mu^*} \Delta_j \mathbb{E}[T_j(n)] \end{aligned}$$

So we can bound the regret simply by bounding each $\mathbb{E}[T_j(n)]$

In fact, we will show that $\mathbb{E}[T_j(n)] \leq \frac{8}{\Delta_j^2} \ln n$ plus a small constant

Detailed Proof of Theorem 1

Number of times
 j has been taken after
 n plays


$$T_j(n) = 1 + \sum_{t=K+1}^n \mathbb{I}\{I_t = j\} \quad (1)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\{I_t = j, T_j(t-1) \geq \ell\} \quad (2)$$

true for any positive integer ℓ

Detailed Proof of Theorem 1

$$T_j(n) = 1 + \sum_{t=K+1}^n \mathbb{I}\{I_t = j\} \quad (1)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\{I_t = j, T_j(t-1) \geq \ell\} \quad (2)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\{U_j(t-1) \geq U^*(t-1), T_j(t-1) \geq \ell\} \quad (3)$$

where at each round n we denote the upper confidence bound:

$$\begin{aligned} U_j(n) &= \bar{x}_j + \sqrt{\frac{2 \ln n}{T_j(n)}} \\ &= \bar{X}_{j, T_j(n)} + c_{n, T_j(n)} \text{ with } c_{n, s} = \sqrt{(2 \ln n)/s} \end{aligned}$$

Detailed Proof of Theorem 1

Relax the event: $\{U_j(t-1) \geq U^*(t-1) \text{ and } T_j(t-1) \geq \ell\}$

UCB of arm j exceeds that of optimal arm \Rightarrow max UCB of arm j during first ℓ trials exceeds minimum UCB of the optimal arm, so:

$$T_j(n) \leq \ell + \sum_{t=K+1}^n \mathbb{I}\{U_j(t-1) \geq U^*(t-1), T_j(t-1) \geq \ell\} \quad (3)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\left\{\max_{\ell \leq s < t} \bar{X}_{j,s} + c_{t-1,s} \geq \min_{0 < s' < t} \bar{X}_{s'}^* + c_{t-1,s'}\right\} \quad (4)$$

Detailed Proof of Theorem 1

Further relaxing event in eq (4): $\{\max_{\ell \leq s < t} \bar{X}_{j,s} + c_{t-1,s} \geq \min_{0 < s' < t} \bar{X}_{s'}^* + c_{t-1,s'}\}$

At least one pair s, s' for which the values of the quantities inside the max/min will satisfy the inequality. So:

$$T_j(n) \leq \ell + \sum_{t=K+1}^n \mathbb{I}\{\max_{\ell \leq s < t} \bar{X}_{j,s} + c_{t-1,s} \geq \min_{0 < s' < t} \bar{X}_{s'}^* + c_{t-1,s'}\} \quad (4)$$

$$\leq \ell + \sum_{t=K+1}^n \sum_{s=\ell}^{t-1} \sum_{s'=1}^{t-1} \mathbb{I}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\} \quad (5)$$

$$\leq \ell + \sum_{t=1}^{\infty} \sum_{s=\ell}^{t-1} \sum_{s'=1}^{t-1} \mathbb{I}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\} \quad (6)$$

Detailed Proof of Theorem 1

To summarize what we have so far:

$$T_j(n) = 1 + \sum_{t=K+1}^n \mathbb{I}\{I_t = j\} \quad (1)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\{I_t = j, T_j(t-1) \geq \ell\} \quad (2)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\{U_j(t-1) \geq U^*(t-1), T_j(t-1) \geq \ell\} \quad (3)$$

$$\leq \ell + \sum_{t=K+1}^n \mathbb{I}\left\{\max_{\ell \leq s < t} \bar{X}_{j,s} + c_{t-1,s} \geq \min_{0 < s' < t} \bar{X}_{s'}^* + c_{t-1,s'}\right\} \quad (4)$$

$$\leq \ell + \sum_{t=1}^{\infty} \sum_{s=\ell}^{t-1} \sum_{s'=1}^{t-1} \mathbb{I}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\} \quad (5)$$

Detailed Proof of Theorem 1

The ♥ of the argument: bound $\mathbb{P}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\}$

➤ Consider three events: $\bar{X}_{s'}^* \leq \mu^* - c_{t,s'}$ (7)

$$\bar{X}_{j,s} \geq \mu_j + c_{t,s} \quad (8)$$

$$\mu^* < \mu_j + 2c_{t,s} \quad (9)$$

Claim: event $\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\}$ implies

- a. One of the 3 events (7),(8),(9) must occur
- b. Event (9) cannot occur with well chosen ℓ
- c. Probability of (7) and (8) occurring can be bounded by Chernoff-Hoeffding

Proof of (a): assume (7) and (8) are false, then (9) is true since

$$\mu_j + 2c_{t,s} > \bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'} > \mu^*$$

Detailed Proof of Theorem 1

Claim (b): $\mu^* \geq \mu_j + 2c_{t,s}$ with well chosen ℓ

Proof of (b): $\mu^* \geq \mu_j + 2c_{t,s}$ for $\ell = \lceil \frac{8 \ln n}{\Delta_j^2} \rceil$

Indeed,

$$\mu^* - \mu_i - 2c_{t,s} = \mu^* - \mu_i - 2\sqrt{\frac{2 \ln t}{s}} \geq \mu^* - \mu_j - \Delta_j = 0$$

$$\text{for } s \geq \frac{8 \ln n}{\Delta_j^2}$$

Detailed Proof of Theorem 1

Recall Chernoff-Hoeffding:

X_1, \dots, X_n are independent random variables with $[0, 1]$ support

Let $\bar{X} = \frac{1}{n} \sum_i X_i$ and $\mu = \mathbb{E}[\bar{X}]$

Then we have:

$$\mathbb{P}(\bar{X} + c < \mu) \leq e^{-2nc^2} \text{ and } \mathbb{P}(\bar{X} - c > \mu) \leq e^{-2nc^2}$$

Yielding (c):

$$\text{Probability of (7)} = \mathbb{P}\{\bar{X}_{s'}^* \leq \mu^* - c_{t,s'}\} \leq e^{-4 \ln t} = t^{-4}$$

$$\text{Probability of (8)} = \mathbb{P}\{\bar{X}_{j,s} \geq \mu_j + c_{t,s}\} \leq e^{-4 \ln t} = t^{-4}$$

Detailed Proof of Theorem 1

Putting it all together using union bound and $\ell = \lceil \frac{8 \ln n}{\Delta_j^2} \rceil$, recall eq (5):

$$T_j(n) \leq \ell + \sum_{t=1}^{\infty} \sum_{s=\ell}^{t-1} \sum_{s'=1}^{t-1} \mathbb{I}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\}$$

So,

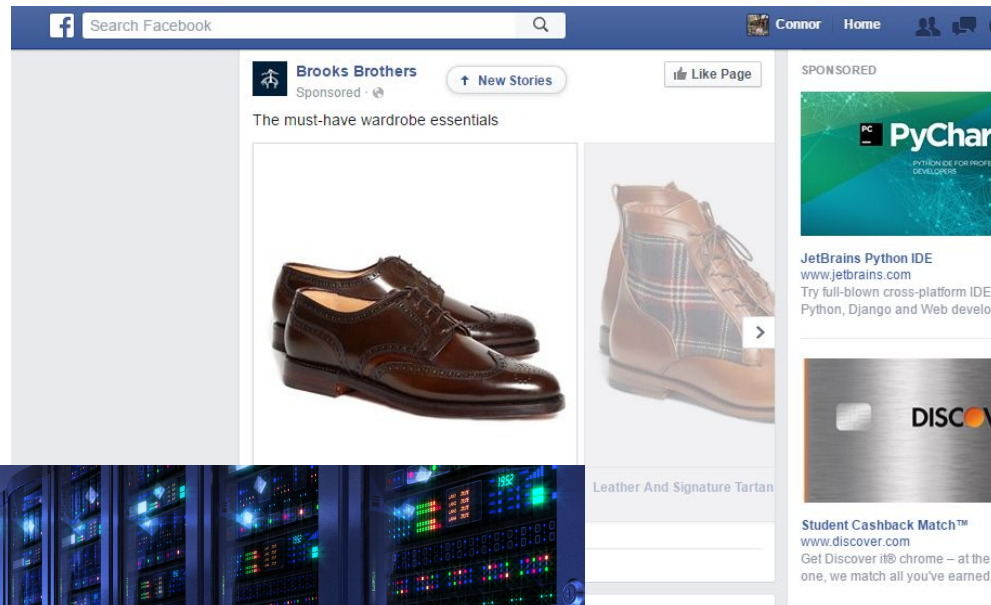
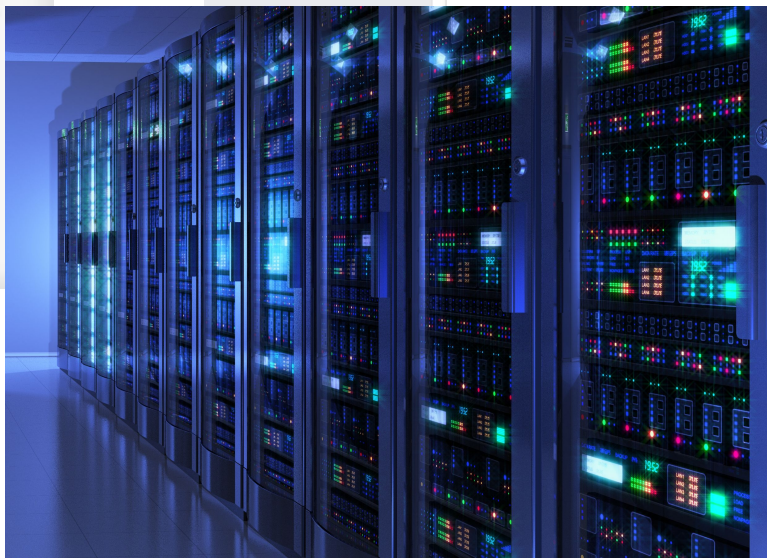
$$\mathbb{E}[T_j(n)] \leq \ell + \sum_{t=1}^{\infty} \sum_{s=\ell}^{t-1} \sum_{s'=1}^{t-1} \mathbb{P}\{\bar{X}_{j,s} + c_{t,s} \geq \bar{X}_{s'}^* + c_{t,s'}\} \quad (10)$$

$$\leq \lceil \frac{8 \ln n}{\Delta_j^2} \rceil + \sum_{t=1}^{\infty} \sum_{s=\lceil \frac{8 \ln n}{\Delta_j^2} \rceil}^{t-1} \sum_{s'=1}^{t-1} \mathbb{P}\{eq(7)\} + \mathbb{P}\{eq(8)\} \quad (11)$$

$$\leq \lceil \frac{8 \ln n}{\Delta_j^2} \rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s'=1}^{t-1} 2t^{-4} \quad (12)$$

$$\leq \frac{8 \ln n}{\Delta_j^2} + 1 + \frac{\pi^2}{3}$$

Real-World Applications



Advertisement Placement

Goal: Place ad that a user will most likely click when a page is rendered.

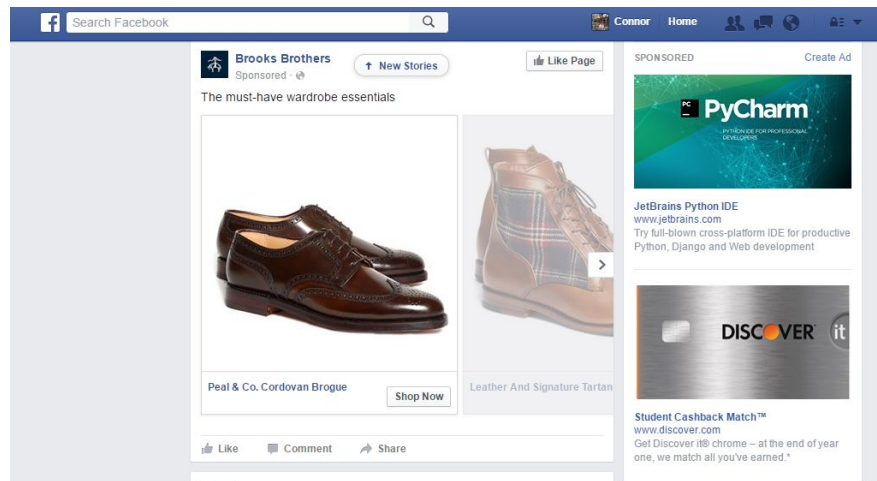
Problem: Don't know if user will click on ad unless it is shown.

Multi-armed bandits solution:

T = sequence of user viewing ads

X = possible ads

c_t = click (1) or no click (0) for x_t



Ethical Clinical Trials

Goal: Find the most effective treatment over time.

Naive solution: Each group of subjects gets a unique treatment.

Problem arises when **new** subjects get bad treatment.



Ethical Clinical Trials

Reduce to multi-armed bandits problem. One of the first motivations for studying multi-armed bandits!

Assume sequential order of subjects.

T = sequence of subjects

X = treatments

c_t = result of treatment x_t



Network Server Selection

Goal: Choose server in distributed system with minimal response time.

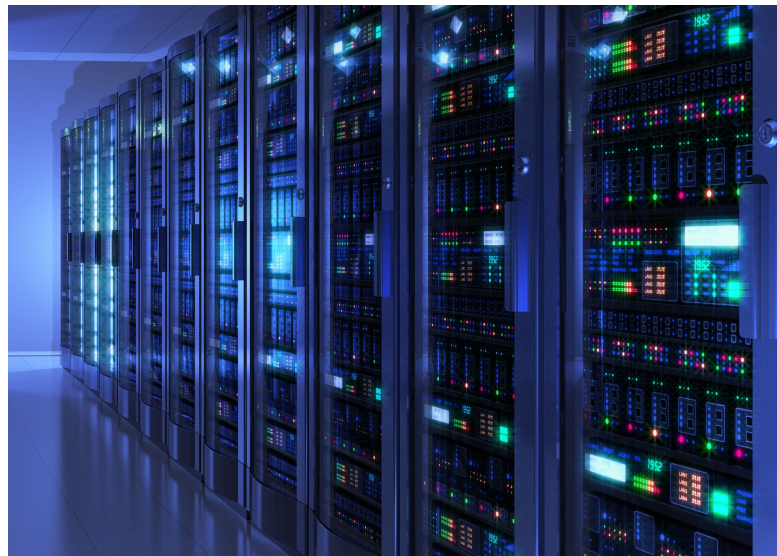
Problem: Don't know server latency until actual connection.

Multi-armed bandits solution:

T = sequence of connections

X = servers

c_t = response time for x_t



Another Exploration Strategy: ϵ -greedy

➤ Simple key idea:

- Pick a parameter $0 < \epsilon < 1$
- At each round
 - Greedily play the arm with highest empirical mean w.p $1 - \epsilon$
 - Play random arm with probability ϵ

➤ Theoretical Result (theorem 3):

$$\text{For } \epsilon_n = \frac{K}{d^2 n}, \text{ regret} = O\left(\frac{K \ln n}{d^2}\right), \text{ provided } 0 < d < \min_{i \neq i^*} \Delta_i$$

➤ Draw-backs:

- Naive exploration for $K > 2$: no distinction of sub-optimal arms
- Requires knowledge of Δ
- Outperformed by UCB in practice

Extensions of UCB1

Monte Carlo Tree Search / UCT - used in first iteration of Go AI

LinUCB - contextual bandits with linear reward functions

UCBogram - contextual bandits with non-linear reward functions

NeuralBandit - using neural nets

UCB-ALP - contextual bandits, used a lot in practice

Adversarial Bandits and EXP3

Adversarial Bandit Setting

➤ Payoff generation process

- No statistical assumption made, can be adversarial
- Similar to previous adversarial online learning setting, payoff cannot depend on the random choices made by the player during the game

➤ Measure of success

- Focus is typically on weak-regret, which measures the regret for the best single action

EXP3 Algorithm

For **non-stochastic** bandits (eg. adversarial bandits)

Similar idea to **multiplicative weights** algorithms

Upper bound of the weak regret:

Theorem 3.1 *For any $K > 0$ and for any $\gamma \in (0, 1]$,*

$$G_{\max} - \mathbf{E}[G_{\mathbf{Exp3}}] \leq (e - 1)\gamma G_{\max} + \frac{K \ln K}{\gamma}$$

holds for any assignment of rewards and for any $T > 0$.

EXP3 Algorithm

Algorithm Exp3

Parameters: Real $\gamma \in (0, 1]$

Initialization: $w_i(1) = 1$ for $i = 1, \dots, K$.

For each $t = 1, 2, \dots$

1. Set

$$p_i(t) = (1 - \gamma) \frac{w_i(t)}{\sum_{j=1}^K w_j(t)} + \frac{\gamma}{K} \quad i = 1, \dots, K.$$

2. Draw i_t randomly accordingly to the probabilities $p_1(t), \dots, p_K(t)$.

3. Receive reward $x_{i_t}(t) \in [0, 1]$.

4. For $j = 1, \dots, K$ set

$$\begin{aligned} \hat{x}_j(t) &= \begin{cases} x_j(t)/p_j(t) & \text{if } j = i_t \\ 0 & \text{otherwise,} \end{cases} \\ w_j(t+1) &= w_j(t) \exp(\gamma \hat{x}_j(t)/K) . \end{aligned}$$

EXP3 vs UCB1

Algorithm Exp3

Parameters: Real $\gamma \in (0, 1]$

Initialization: $w_i(1) = 1$ for $i = 1, \dots, K$.

For each $t = 1, 2, \dots$

1. Set

$$p_i(t) = (1 - \gamma) \frac{w_i(t)}{\sum_{j=1}^K w_j(t)} + \frac{\gamma}{K} \quad i = 1, \dots, K.$$

2. Draw i_t randomly accordingly to the probabilities $p_1(t), \dots, p_K(t)$.

3. Receive reward $x_{i_t}(t) \in [0, 1]$.

4. For $j = 1, \dots, K$ set

$$\begin{aligned} \hat{x}_j(t) &= \begin{cases} x_j(t)/p_j(t) & \text{if } j = i_t \\ 0 & \text{otherwise,} \end{cases} \\ w_j(t+1) &= w_j(t) \exp(\gamma \hat{x}_j(t)/K) . \end{aligned}$$

Deterministic policy: UCB1.

Initialization: Play each machine once.

Loop:

- Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where \bar{x}_j is the average reward obtained from machine j , n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

Sources

<http://homes.di.unimi.it/~cesabian/Pubblicazioni/ml-02.pdf>

<http://arxiv.org/pdf/1204.5721v2.pdf>

<http://www.cs.cornell.edu/courses/cs683/2007sp/lecnotes/week8.pdf>

<http://jeremykun.com/2013/10/28/optimism-in-the-face-of-uncertainty-the-ucb1-algorithm/>

<http://cseweb.ucsd.edu/~yfreund/papers/bandits.pdf>