Group 4

Multi Armed Bandits

Pranav | Manideep | Sri Charan

TABLE OF CONTENTS

1

THE PROBLEM STATEMENT

What is this Multi Armed Bandit?

Explore-Exploit Dilemma

- You're in Vegas, faced with a row of slot machines
- No idea which machines readily give rewards and which are real stingy.
- After a few pulls, you have a rough guess which machine is looking profitable and which isn't.
- Do you continue to **exploit** your currently best machine, or **explore** other machines: maybe they're gold mines waiting to be discovered.

2

GETTING FORMAL

Let’s kick in some math. Shall we?

Problem Formulation

- Row of 'k' slot machines(strategies) - $1, 2, \dots, k$
- Reward for playing strategy s is a r.v. θ_s with distribution π_s, mean μ_s
- At each iteration 't'
  - slot $s(t)$ is played according to algorithm ALG
  - Expected reward for this step = $\mathbb{E}_{\theta_{s(t)}}[\theta_{s(t)}] = \mu_{s(t)}$
- Best strategy, $s^* = \operatorname{argmax}_{i \in [k]} \mu_i$
- Regret $R(n, ALG)$ over n iterations

$$R(n, ALG) = B - \sum_{1 \le t \le n} \mathbb{E}_{\theta_{s(t)}}[\theta_{s(t)}]$$

where B is a regret benchmark

$$= \sum_{1 \le t \le n} \mu_{s^*} - \mu_{s(t)}$$

Impossibility Result

$$R(n, ALG) = \sum_{1 \le t \le n} \mu_{s^*} - \mu_{s(t)} = \sum_{s \in [n] \setminus s^*} (\mu_{s^*} - \mu_s) \mathbb{E}_\theta [T_s(n)]$$

$T_s(n)$: Number of trials of machine s in a sequence of n trials

From Lai and Robbins [1], we have

$$\lim_{n \to \infty} \frac{R(n, ALG)}{\ln n} \ge \sum_{s \in [n] \setminus s^*} \frac{(\mu_{s^*} - \mu_s)}{D_{KL}(\pi_s, \pi_{s^*})}$$

$$\Rightarrow R(n, ALG) = O(\ln n)$$

- Visualization:

$$\mathbb{E}_\theta [T_s(n)] \ge \frac{\ln n}{D_{KL}(\pi_s, \pi_{s^*})}$$

Impossibility Result

$$R(n, ALG) = \sum_{1 \le t \le n} \mu_{s^*} - \mu_{s(t)} = \sum_{s \in [n] \setminus s^*} (\mu_{s^*} - \mu_s) \mathbb{E}_\theta[T_s(n)]$$

$T_s(n)$: Number of trials of machine s in a sequence of n trials

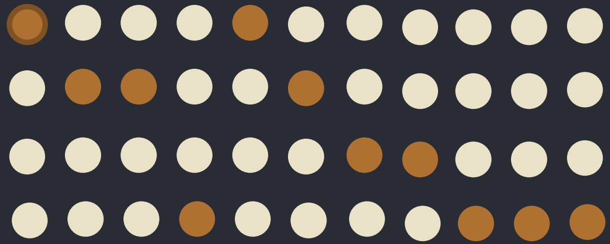
From Lai and Robbins [1], we have

$$\lim_{n \to \infty} \frac{R(n, ALG)}{\ln n} \ge \sum_{s \in [n] \setminus s^*} \frac{(\mu_{s^*} - \mu_s)}{D_{KL}(\pi_s, \pi_{s^*})}$$

$$\Rightarrow R(n, ALG) = O(\ln n)$$

- Visualization:

$$\mathbb{E}_\theta[T_s(n)] \ge \frac{\ln n}{D_{KL}(\pi_s, \pi_{s^*})}$$



Ref -T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6:4–22, 1985.

Impossibility Result

$$R(n, ALG) = \sum_{1 \le t \le n} \mu_{s^*} - \mu_{s(t)} = \sum_{s \in [n] \setminus s^*} (\mu_{s^*} - \mu_s) \mathbb{E}_\theta [T_s(n)]$$

$T_s(n)$: Number of trials of machine s in a sequence of n trials

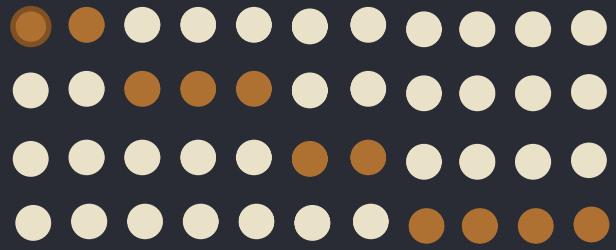
From Lai and Robbins [1], we have

$$\lim_{n \to \infty} \frac{R(n, ALG)}{\ln n} \ge \sum_{s \in [n] \setminus s^*} \frac{(\mu_{s^*} - \mu_s)}{D_{KL}(\pi_s, \pi_{s^*})}$$

$$\Rightarrow R(n, ALG) = O(\ln n)$$

- Visualization:

$$\mathbb{E}_\theta [T_s(n)] \ge \frac{\ln n}{D_{KL}(\pi_s, \pi_{s^*})}$$



Impossibility Result

$$R(n, ALG) = \sum_{1 \le t \le n} \mu_{s^*} - \mu_{s(t)} = \sum_{s \in [n] \setminus s^*} (\mu_{s^*} - \mu_s)\mathbb{E}_\theta[T_s(n)]$$

$T_s(n)$: Number of trials of machine s in a sequence of n trials

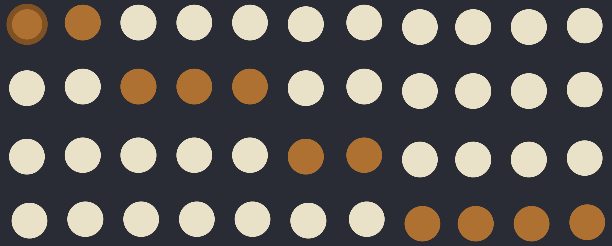
From Lai and Robbins [1], we have

$$\lim_{n \to \infty} \frac{R(n, ALG)}{\ln n} \ge \sum_{s \in [n] \setminus s^*} \frac{(\mu_{s^*} - \mu_s)}{D_{KL}(\pi_s, \pi_{s^*})}$$

$$\Rightarrow R(n, ALG) = O(\ln n)$$

- Visualization:

$$\mathbb{E}_\theta[T_s(n)] \ge \frac{\ln n}{D_{KL}(\pi_s, \pi_{s^*})}$$



Biased coin:

$$\mathbb{P}(H) = 0.5 - \epsilon$$

Unbiased coin:

$$\mathbb{P}(H) = 0.5$$

T the number of tosses required to determine with $l(= say\ 0.99)$ surety

$$T \ge O(\frac{1}{\epsilon^2})$$

Ref: https://people.cs.umass.edu/~arya/courses/650/lecture11.pdf

3

GAMBLING STRATEGIES

How to get rich playing a multi armed bandit.

1. Epsilon - Greedy Strategy

1. We keep track of the average reward we've earned from each arm, as the trials progress.

1. Choose the current "best arm" with probability $1-\varepsilon$ and the others each with a probability $\varepsilon/n-1$

1. Update the average reward of the arm chosen based on the outcome

1. $P(a(t) \neq a^*) = \varepsilon$ as $t \rightarrow \infty$

1. Expected Cumulative Regret is of order $O(n)$

2. This algorithm basically says explore with probability ε and exploit with probability $1-\varepsilon$

2. Epsilon-decreasing Strategy

1. Similar to Epsilon-greedy strategy but the probability of choosing the current best arm is a **decreasing function** of t

1. One such function for epsilon is $\varepsilon(t) = \varepsilon_0/t$

1. The exploration phase vanishes after sufficient number of trials

1. $P(a(t) \neq a^*) = \varepsilon_0/t$ as $t \rightarrow \infty$

1. The expected cumulative regret is of logarithmic order i.e O(logn)

1. If we take $\varepsilon(t) = \varepsilon_0/t^2$, then the **exploration phase vanishes much faster**. Too fast for the algorithm to have made a reasonable guess for which is the best arm.

Ref : https://youtu.be/Qy-vum3GH-s

Hoeffding's inequality

$$P(E[\overline{X}] - \overline{X} \geq t) \leq e^{-2nt^2}$$

where n is the number of samples

Principle of Optimism in face of Uncertainty

Certain classes of algorithms for multi armed bandits are based on the principle of optimism in the face of uncertainty. UCB1 is one of the most important algorithm based on it. Suppose you have a black box/oracle that gives you an upper bound on the expected reward for all arms and it gets better each time. This bound reaches the true value after sufficiently large number of trials.

3. Upper Confidence Bound (UCB)

1. There is no oracle that can give the upper bound to the expected reward with certainty

1. There are probabilistic bounds, that can be derived from Hoeffding's inequality

3. It can be seen that

$$\Pr[R(a) - R'(a) > \sqrt{\frac{\log(\frac{1}{\delta})}{2n_a}}] \leq \delta$$

4. Take $\delta = 1/n^4$. The bound becomes $2\log n/n_a$

1. For each arm a , calculate $u_a = R'(a) + \sqrt{\frac{2\log n}{n_a}}$

2. Choose the arm with the largest u_a . Let this arm be

3. Update $n_a = n_a + 1$ for the arm chosen

4. Update $R'(a) = \frac{n_a . R'(a) + reward}{n_a + 1}$

   for the arm chosen

5. Go to step 1

4. Thompson Sampling

- Bayesian approach unlike the previous ones
- Doesn't assume a fixed mean
- The means of each arm are assumed to have some distribution and the distributions get updated based on the rewards
- The arm is chosen based on the value obtained by sampling the distributions
- Usually each arm's mean is assumed to have a beta distribution with parameters x,y different for each arm

1. Initialise x = 1, y= 1 for all arms
1. Now take k samples from each distribution and calculate the average for each of the arms
1. Choose the arm with the highest average and get the reward r
1. Update : x = x+r , y = y + 1-r for the chosen arm
1. Go to step 2

Ref: - https://youtu.be/jlcbEZTgisQ

Does β distribution update to another β-distribution??

Yes!

$$f(p)\alpha p^{x-1}(1-p)^{y-1}$$

$$f(p/r=1)\alpha f(r=1/p).f(p)=p.p^{x-1}(1-p)^{y-1}=p^{x}(1-p)^{y-1}$$

$$f(p/r=0)\alpha f(r=0/p).f(p)=(1-p).p^{x-1}(1-p)^{y-1}=p^{x-1}(1-p)^{y}$$

4

SIMULATIONS

Let’s have a computer back up the bold claims we made “in theory”.

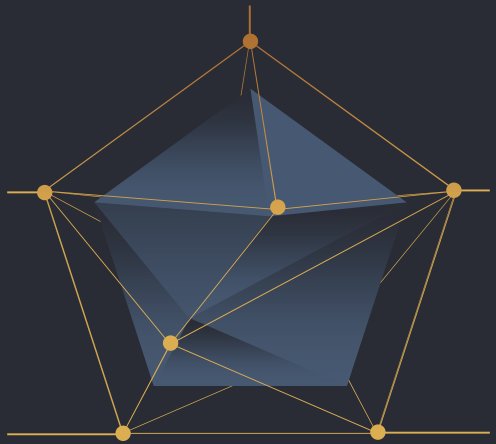
SOLUTIONS WE'LL SIMULATE

EPSILON DECREASE - 2

The value of epsilon decreases as $1/n^2$

EPSILON DECREASE - 1

The value of epsilon decreases as $1/n$



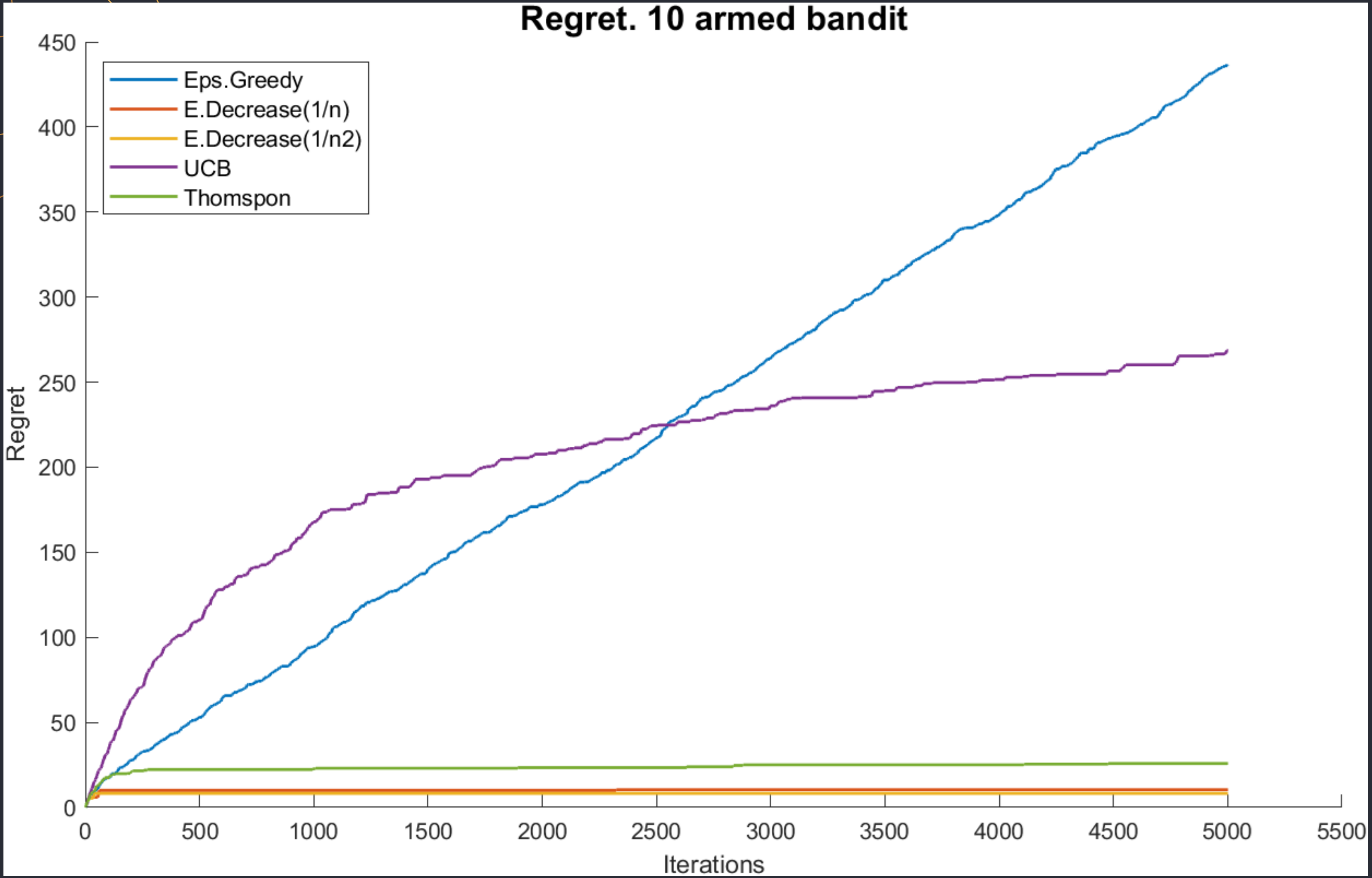
UPPER CONFIDENCE BOUND

EPSILON GREEDY

THOMPSON SAMPLING

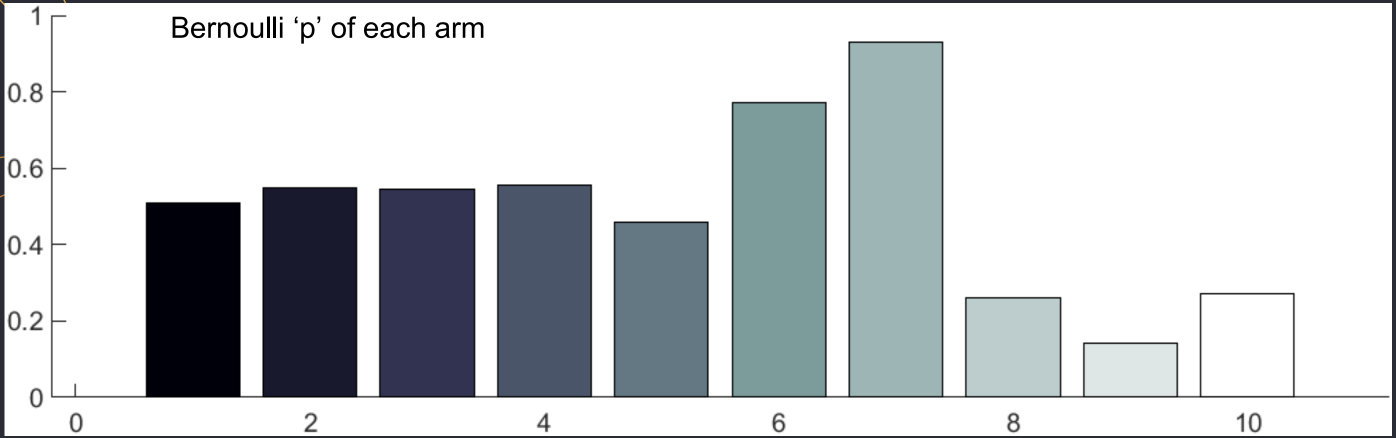
Models the reward of an arm as a Beta distribution

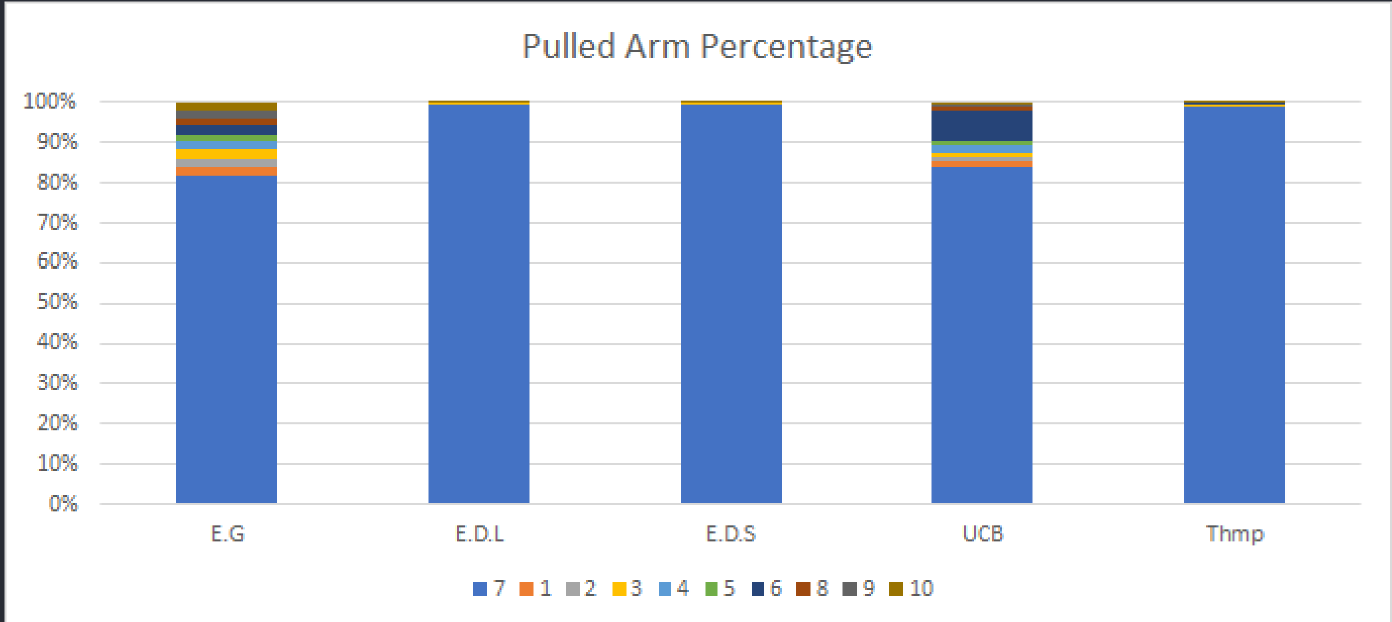
REGRET



Regret. 10 armed bandit
Eps.Greedy
E.Decrease(1/n)
E.Decrease(1/n2)
UCB
Thomspon
Regret
Iterations
0
50
100
150
200
250
300
350
400
450
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500

NUMBER OF PULLS



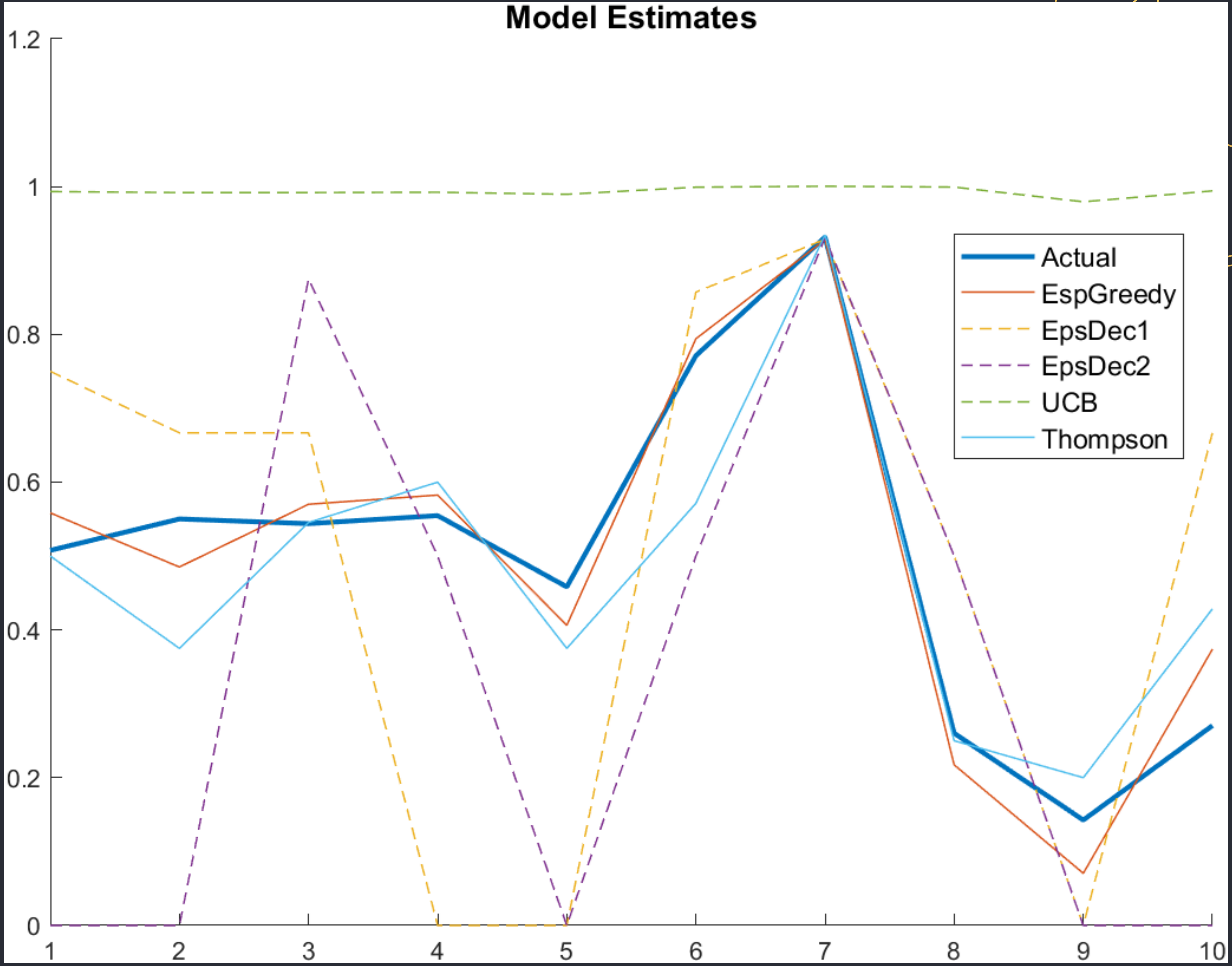


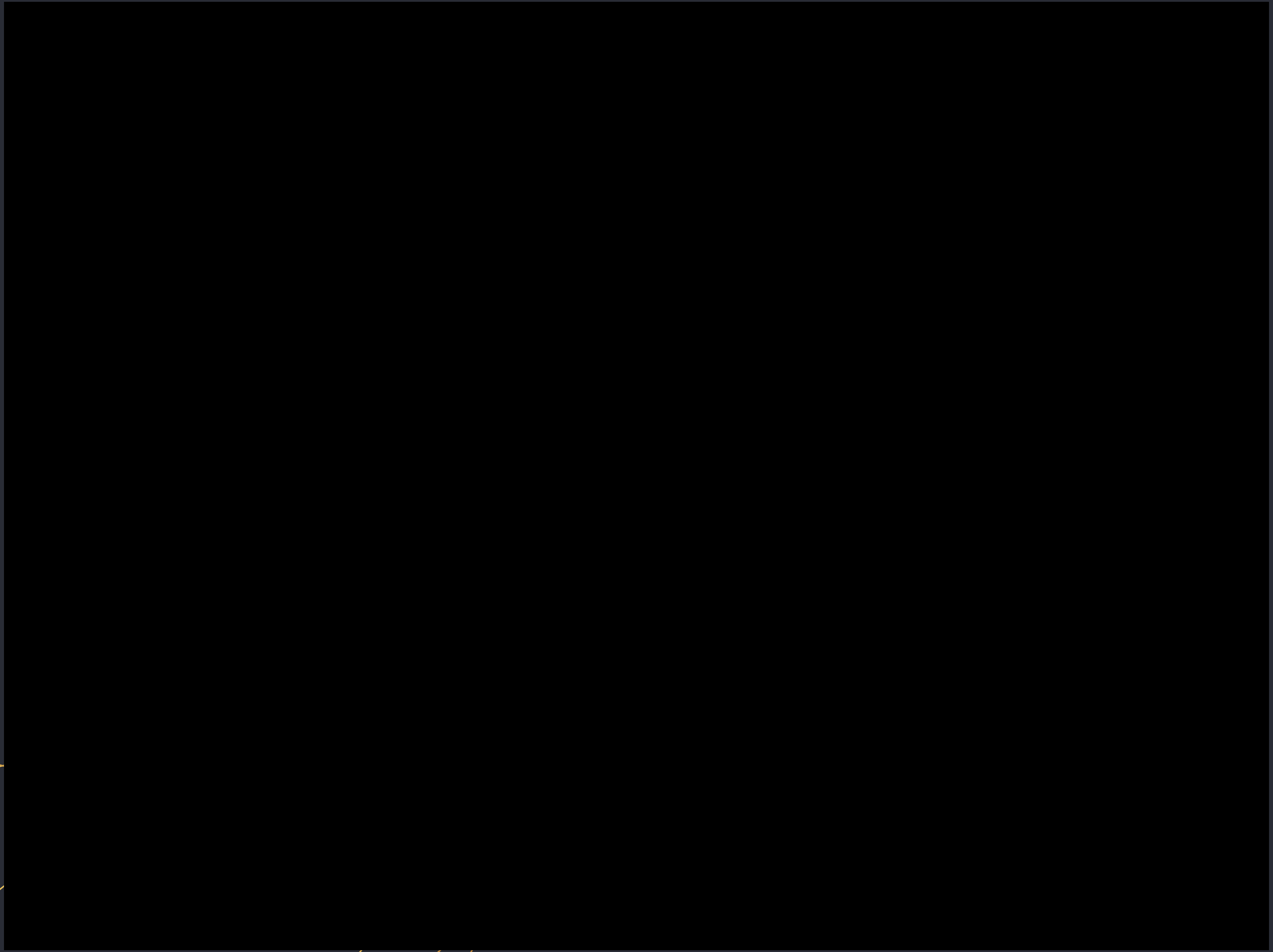
ESTIMATING ARM PROBABILITIES

Eps Greedy and Thompson have the best estimates.

UCB is an upper bound anyway, it's not supposed to model it well.

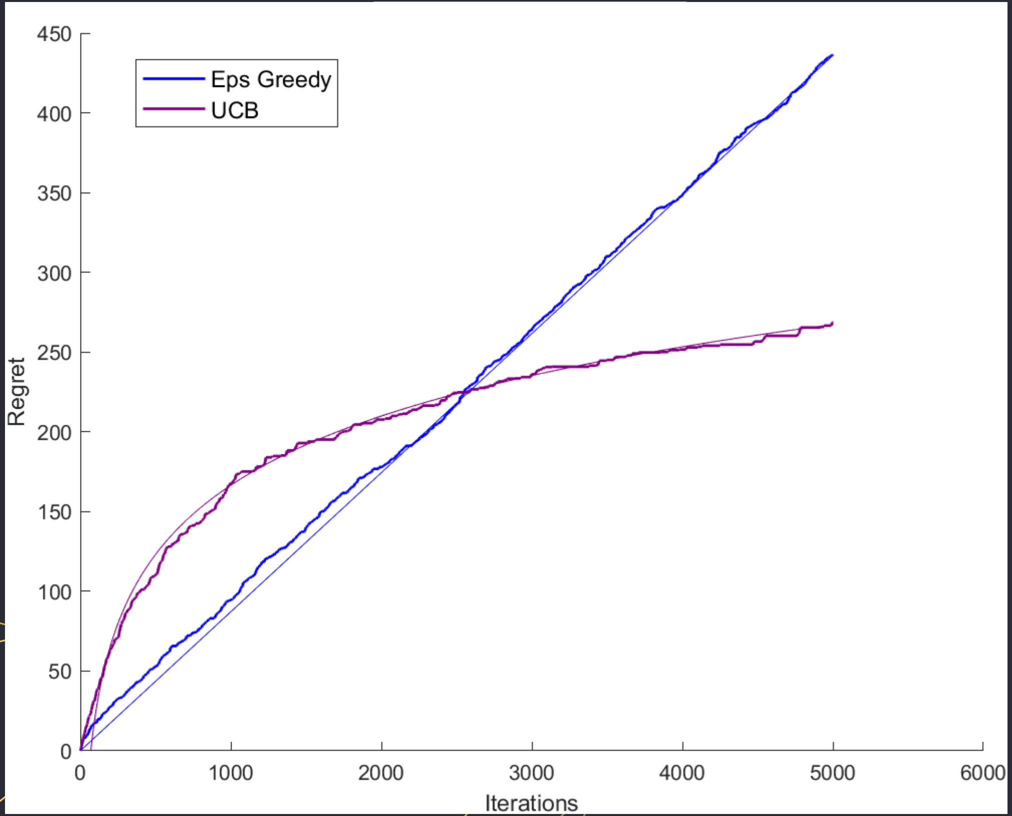
Though Eps Decay strategies model the distribution as a whole poorly, **every model has correctly estimated the probability of the best arm.**





THOMPSON SAMPLING

THEORETIC COMPARISON



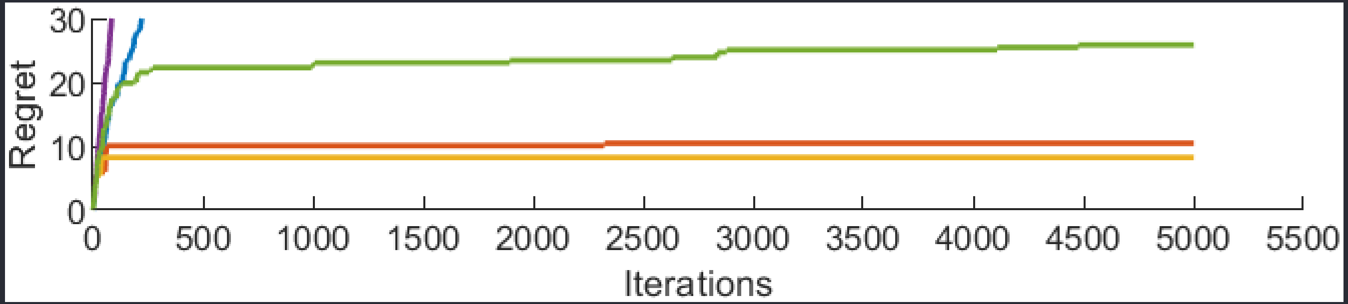
Very satisfyingly, the results agree almost too well with the theory.

	THEORETIC	MODELLED
EPS GREEDY	O(n)	0.087*n
UCB	O(log.n)	62.6*log(n) - 266

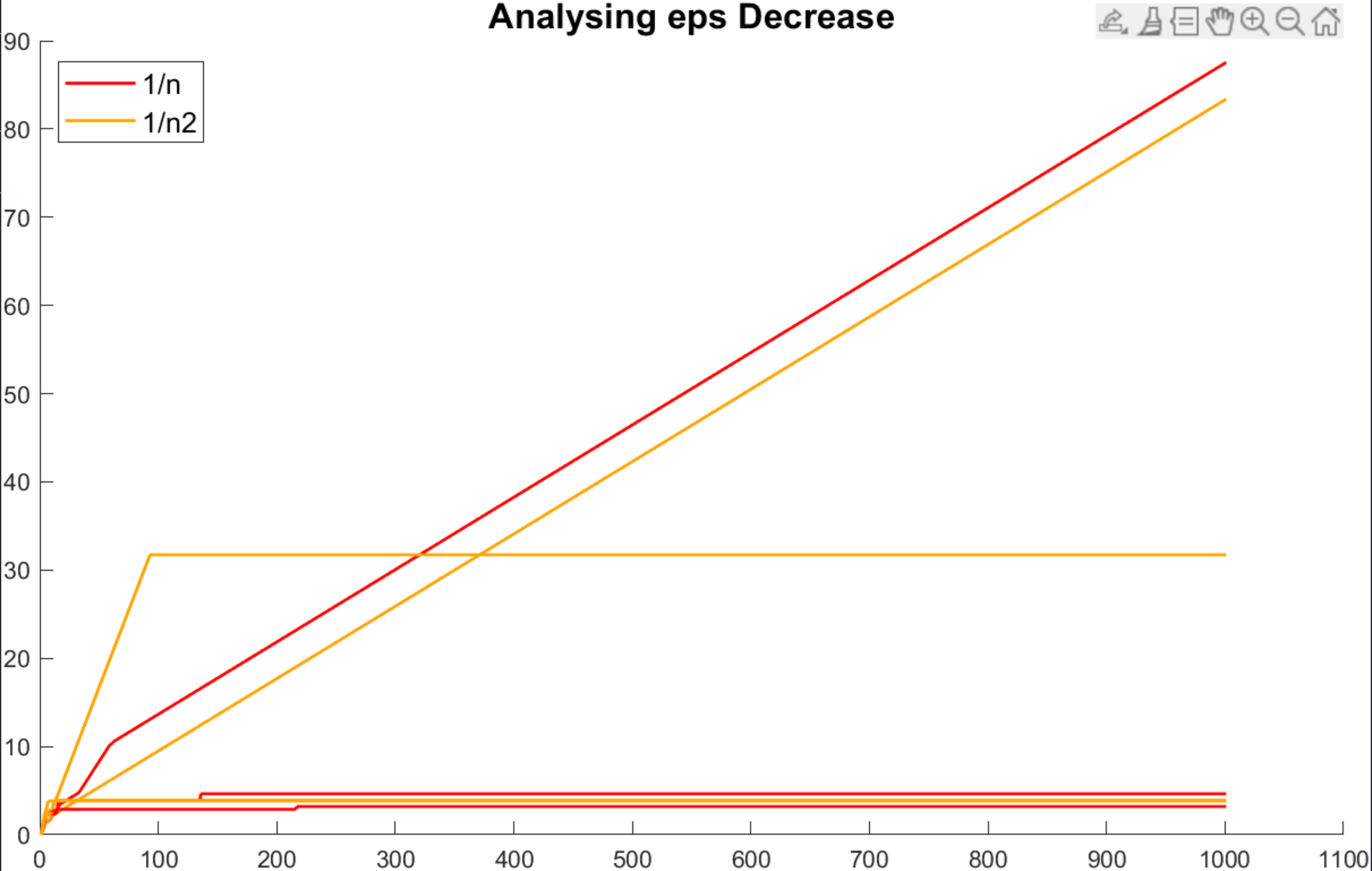
WHAT'S THE DEAL WITH THE EPSILON DECREASE STRATEGIES?

Why on earth do simple epsilon strategies perform so well? These eps. **Decrease** strategies stop exploring pretty quick as epsilon decays to zero. After a while, all they do is exploit. And they perform great. ***If*** they've managed to find the right arm, that is.

The problem is, **the faster you decay, the less time you have to find the right arm.** We shall see this in full force in the next slide.



EPS DECREASE

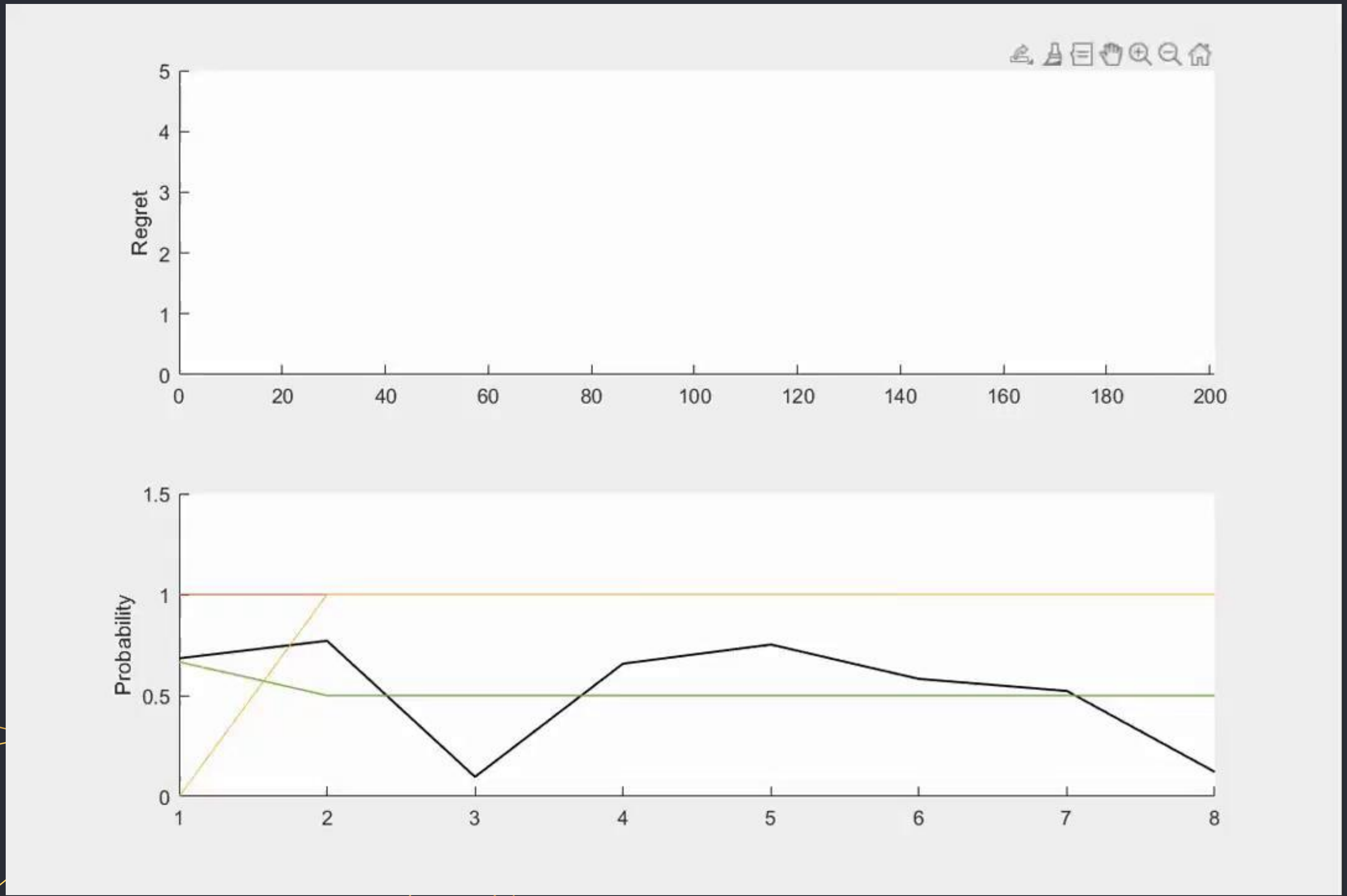


NON-STATIONARY BANDITS

As we have seen, the epsilon decrease strategies have been absolutely destroying our MAB (on their good days). Let’s make things a little more challenging.

A non-stationary bandit is one that randomly changes the Bernoulli random variable ‘p’ for each arm. For simplicity, this random change will happen after every 200 iterations.

When the eps decrease strategies cannot now “cling onto their best arm”, let us see what becomes of them. And of the rest.



NON-STATIONARY SIMULATION

5

VARIANTS OF M.A.B.

As if the original wasn't bad enough.

ADVERSARIAL BANDIT

- An adversary can control the reward function

COMBINATORIAL BANDIT

- Non linear reward function
- Example : No reward when you win three consecutive turns

DUELLING BANDIT

- Pitching two strategies against each other.
- Play two slot machines, you only know if one is performing better than the other.

6

REAL WORLD APPLICATIONS

ETHICAL CLINICAL TRIALS

- In a treatment of disease
- Evaluation of k treatments on n patients.
- Reward cures and penalize deaths.

INTERNET AD PLACEMENT

- k Advertisers
- n Users visiting the website
- Reward if they click it.

SERVER SELECTION IN NETWORKS

- Client looks for k servers
- For n instances try different servers
- Cost is the time delay

PRICING IDENTICAL GOODS FOR SALE

- k Pricing strategies
- n Markets
- Reward is the profit made.

Ref:- http://www.cs.cornell.edu/courses/cs683/2007sp/lecnotes/week8.pdf

Thank you!