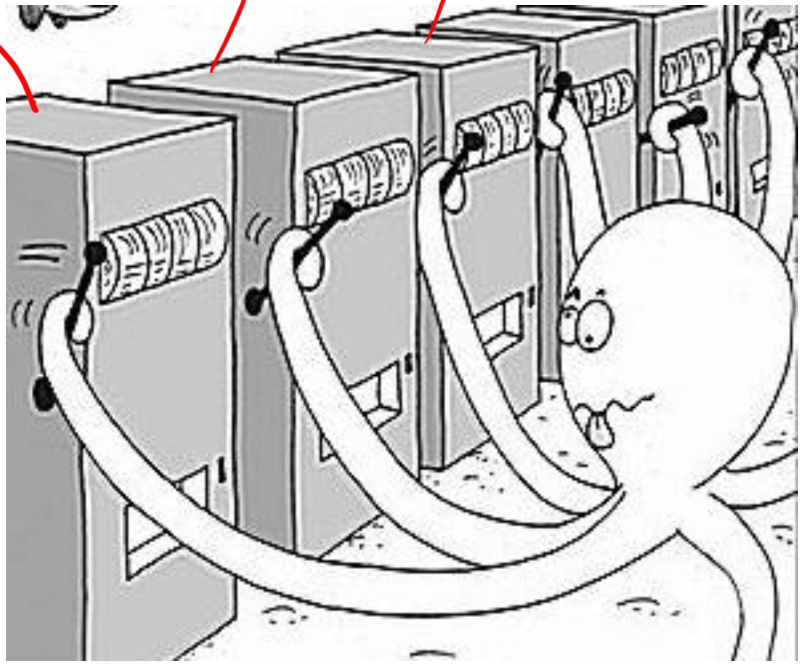
Multi-Armed Bandits

Daniel Brown

r -2 0.5 +1

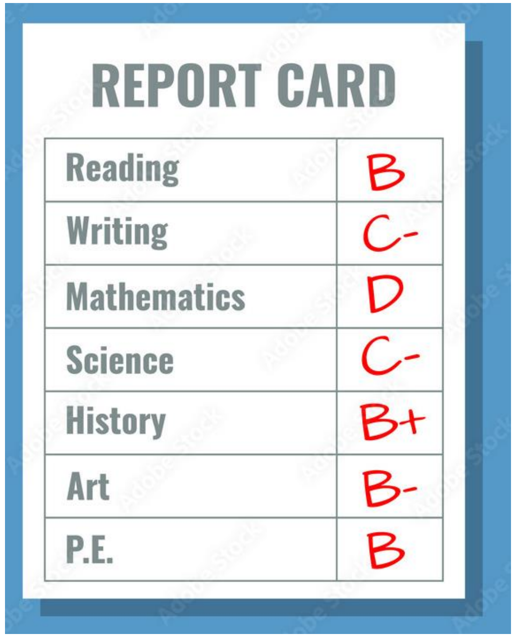


Evaluative feedback

Not supervised learning
– We're not given labels

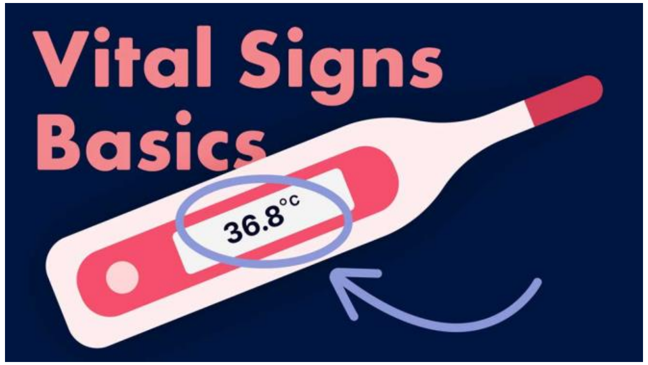
+1

-1



REPORT CARD

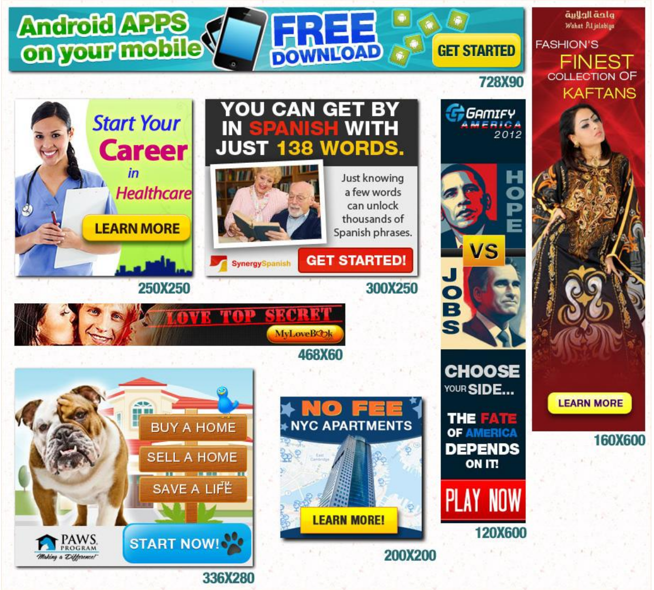
Reading	B
Writing	C-
Mathematics	D
Science	C-
History	B+
Art	B-
P.E.	B



MAB – reward
– action space

Applications

- Online Advertising and Recommendation
- Clinical Trials
- Robotics
- Dynamic Pricing
- Search Engine Optimization
- Education and Learning Platforms



$\pi \rightarrow \arg\max_i \mu_i$ $\mu_1 = \mathbb{E}[r(a_1)]$

$r_t(a_1) \sim N(0,1) \quad \forall t$

$r(a_2) \sim U[-2,2]$

$r(a_3) = N(10,100)$

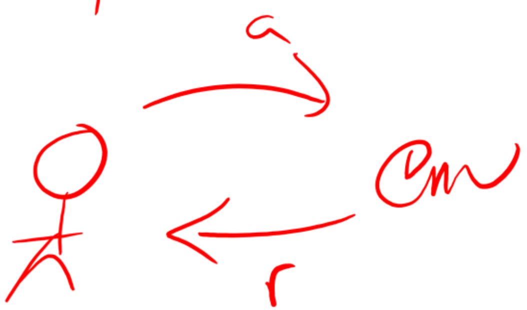
Problem formalism

Actions

- Arms $\mathcal{A} = \{a_1, \ldots, a_k\}$
  - Each arm is associated with an unknown reward distribution
- Rewards $r_t(a_i)$
- Possible Goals
  - ★ Maximize cumulative reward (Minimize regret)
  - Best arm identification
- Standard Assumptions
  - Independence: Rewards from each arm are independent
  - Stationarity: Reward distributions don't change over time

stochasticity is allowed

predone



1-step interaction

How should we solve this problem?

1) initial exploration of actions

- figure out which has highest mean reward and take that action

median, std, full dist.

"Smart" exploration

Exploitation

Balance exploration & exploitation

Random

- maybe good at first

pure exploration

bad exploitation

Greedy

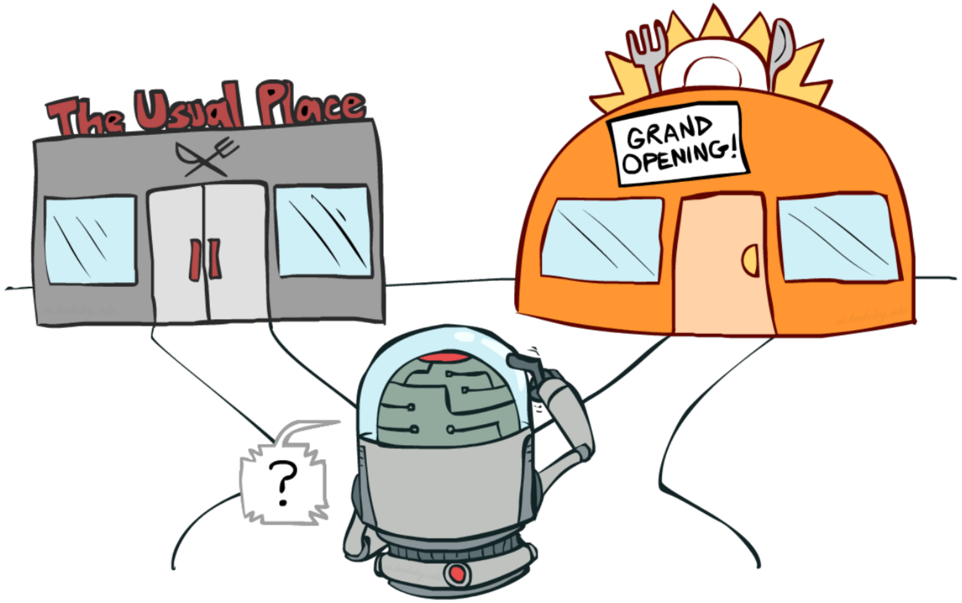
$\bar{\mu}_i$ = sample ave

	a_1		a_2		a_3
$t=1$	0	$t=2$	0	$t=3$	1
$t=4$					-1

$\bar{\mu}_1 = 0 \qquad \bar{\mu}_2 = 0 \qquad \bar{\mu}_3 = 0$

Not Good

Exploration



ϵ-Greedy

$\varepsilon \in [0,1]$ determines the prob of taking a rand action

pick ε

randomly generate $x \in [0,1]$

if $x \leq \varepsilon$

take rand action uniformly

else

take greedy action

★ Anneal ε : start $\varepsilon = 1$ then $\varepsilon \to 0$ as $t \to \infty$

2.3 The 10-armed Testbed

To roughly assess the relative effectiveness of the greedy and ε-greedy action-value methods, we compared them numerically on a suite of test problems. This was a set of 2000 randomly generated k-armed bandit problems with $k = 10$. For each bandit problem, such as the one shown in Figure 2.1, the action values, $q_*(a)$, $a = 1, \ldots, 10$,

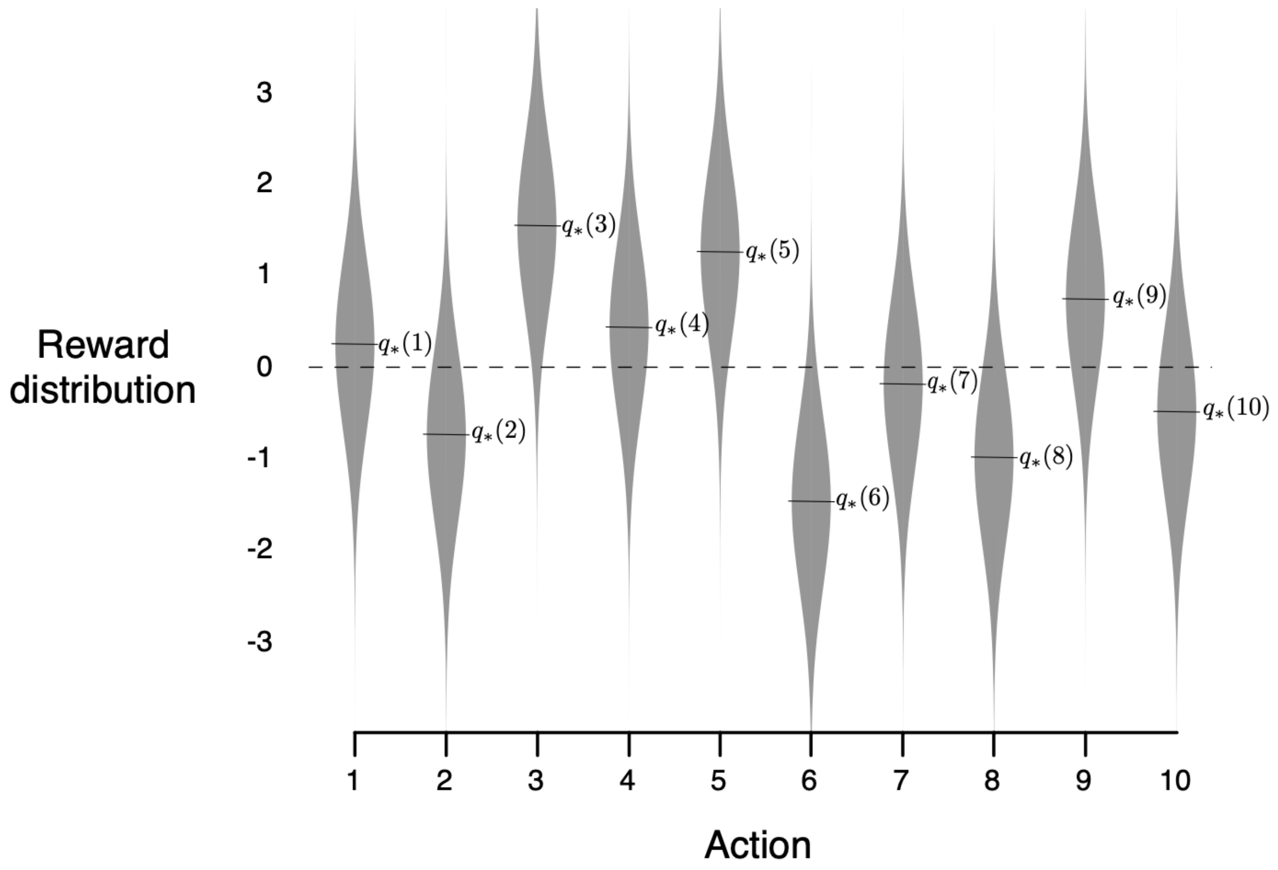
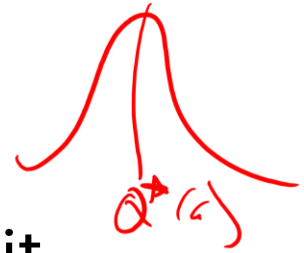


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.

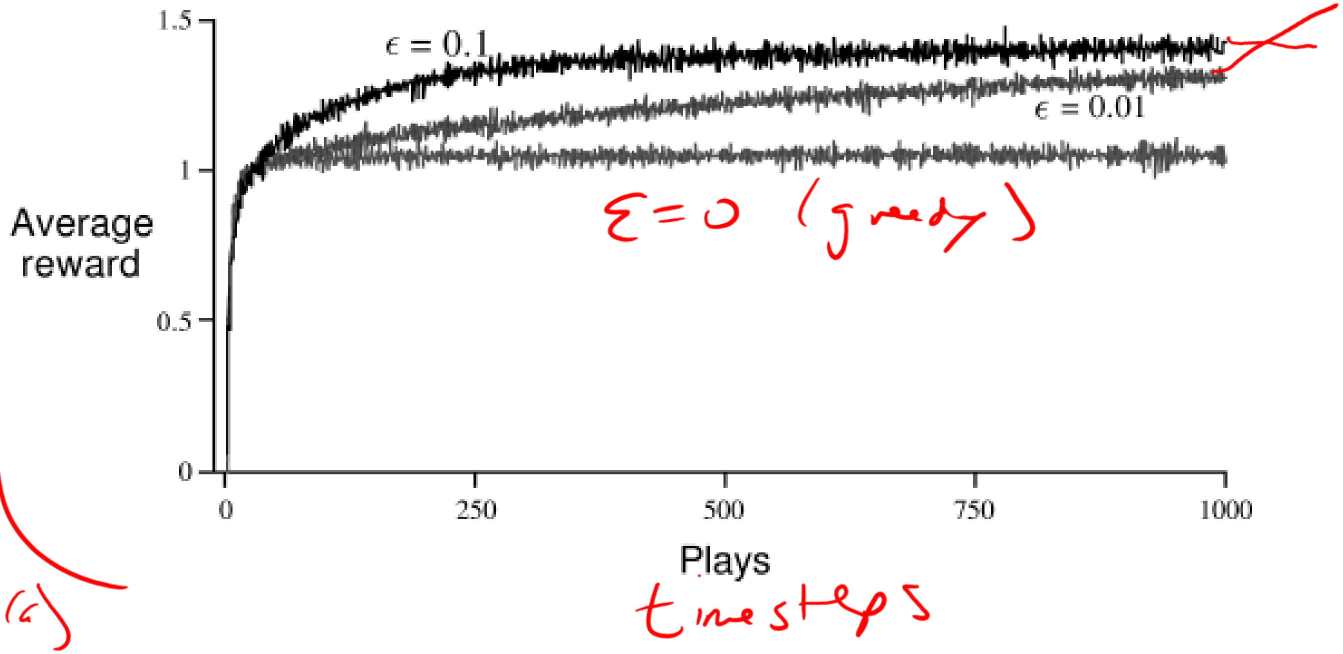
Sutton/Barto figure

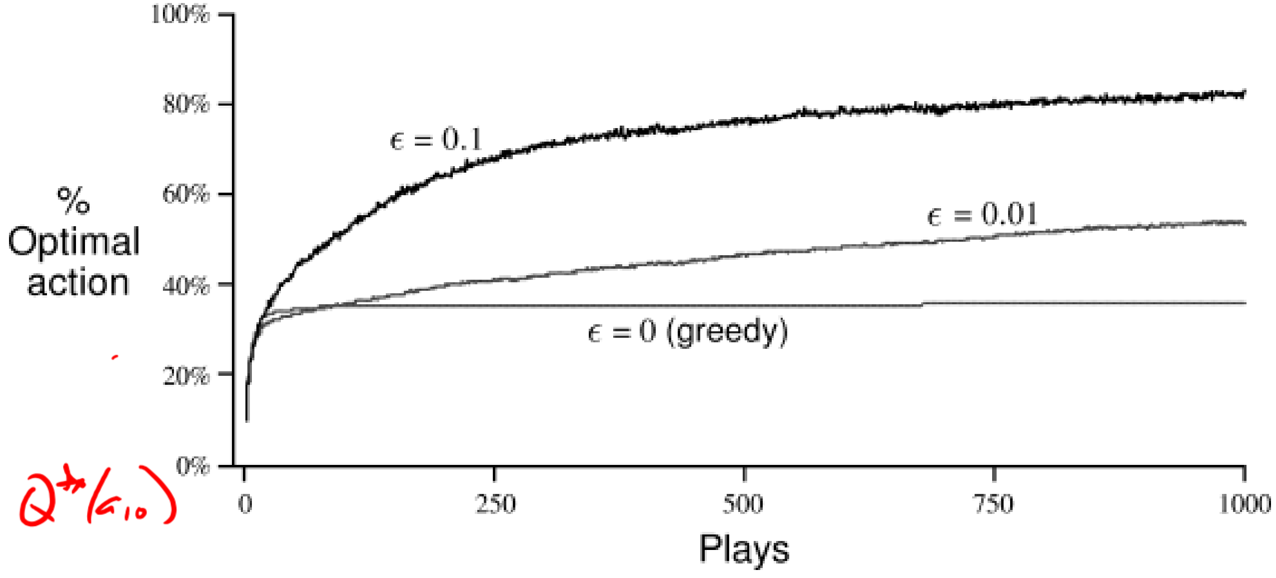
- 10 arms
- Each arm has stochastic reward
  $$r \sim N(Q^*(a), 1)$$
- Averaged over 2000 bandit problems where each problem starts with $Q^*(a) \sim N(0,1)$ for all a



$Q^*(a) = \mathbb{E}[r(a)]$

sample 10 times to get $Q^*(a_1) \dots Q^*(a_{10})$





Problems?

never stops exploring

explores random

Boltzmann (Softmax) Exploration

$$Q(a) = \frac{1}{n}\sum_{i=1}^{n} r_i$$ sample average

$$P(a_i) = \frac{\exp(\beta Q(a_i))}{\sum_a \exp(\beta Q(a))}$$

β = inv. temp

$\beta = 0$
$\Rightarrow$ uniform random

$\beta \to \infty$
$\Rightarrow$ Greedy

$\beta = 1, 10$

Chernoff-Hoeffding Inequality

$0.6 = \bar{X}$

- Let X be a random variable in the range [0,1] and $x_1, x_2, \ldots, x_n$ be n independent and identically distributed samples of X.
- Let $\bar{X} = \frac{1}{n}\sum_i x_i$ (the empirical average)
- Then we have $P(\bar{X} \geq \mathbb{E}[X] + c) \leq e^{-2nc^2}$

$P(\bar{X} \leq \mathbb{E}[X] - c) \leq e^{-2nc^2}$

$\Rightarrow P(|\bar{X} - \mathbb{E}[X]| \geq c) \leq 2e^{-2nc^2}$

$n \to \infty$

$\frac{1}{e^{2nc^2}} \to 0$

$c \to \infty$

$c = 0.4$ $\bar{X} = 0.9$

Some fun math

$n = 10$

$\delta = 0.05$

- $P(\bar{X} \geq \mathbb{E}[X] + c) \leq e^{-2nc^2}$
- Typically, we want to pick some kind of high confidence $1 - \delta$ such that we are very confident about our sample mean being close to the true expectation.
- If we want

$$P(\bar{X} \geq \mathbb{E}[X] + c) \leq \delta$$

$$\boxed{-\log \delta = \log 1/\delta}$$

What is c in terms of δ?

$$\delta = e^{-2nc^2} \Rightarrow \log \delta = -2nc^2 \Rightarrow \sqrt{\frac{\log \delta}{-2n}} = c \Rightarrow c = \sqrt{\frac{1}{2n} \log(1/\delta)}$$

More math

$$c = \sqrt{\frac{1}{2n}\log(1/\delta)}$$

- We can pick δ to be whatever we want, so let's pick
- If we select $\delta = \frac{1}{t^2}$

$\delta \to 0 \quad t \to \infty$

$a \log b = \log b^a$

What is c?

$$c = \sqrt{\frac{1}{2n}\log t^2} = \sqrt{\frac{2}{2n}\log t}$$

$$P(\bar{R} \geq \mathbb{E}[R] + c) \leq 1/t^2$$

$$c = \sqrt{\frac{\log t}{n}}$$

UCB1 (UCB = Upper Confidence Bound)

arms $a_1 \dots a_K$, counts $n_1 \dots n_K$

Key Idea: Optimism in the face of uncertainty

- Play each action once to get initial averages of arm values
- Keep track of counts of pulls for each arm n_i
- At each step t, select $\arg\max_i \bar{X}_i + c(i,t)$
  - Where $c(i,t) = \sqrt{\frac{log(t)}{n_i}}$

$\bar{X}_i$: sample average for arm i

c_1 c_2 c_3

$\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$

a_1 a_2 a_3

$n_i \to \infty$

$c(i,t) \to 0$

$t \to \infty$

Regret

$Q^* = \max_a Q^*(a)$

- Define μ^* as the maximum expected payoff over all k arms
- Regret(T) = $T\mu^* - \sum_{t=1}^{T} r_t$

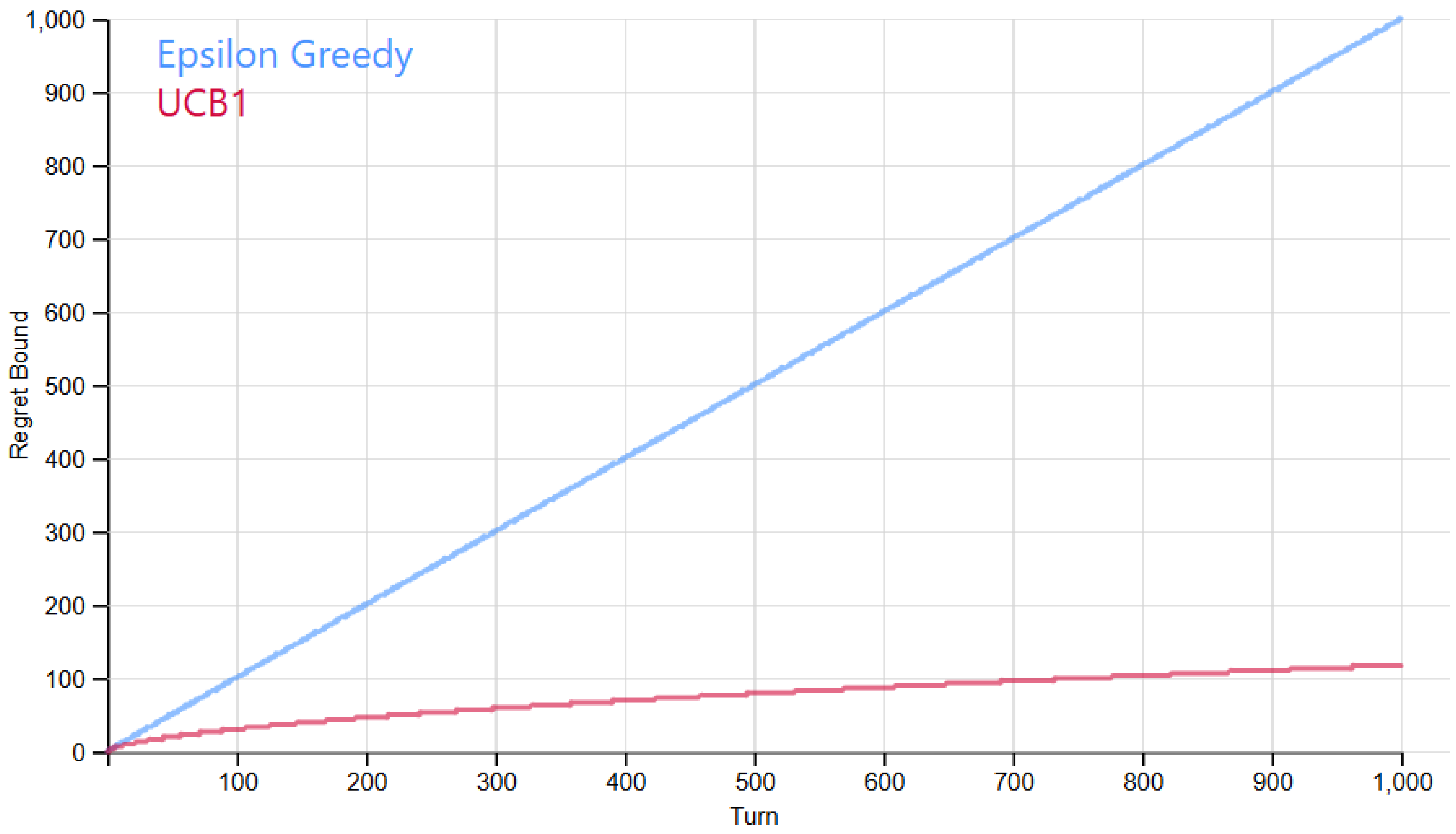
- Epsilon-Greedy Regret
  - $O(T)$
- UCB1 Regret
  - $O(\sqrt{kTlog(T)})$

    k = # arms

T = # timesteps

- A **No-Regret** algorithm is such that Regret(T)/T → 0 as $T \to \infty$
  - Average regret goes to zero

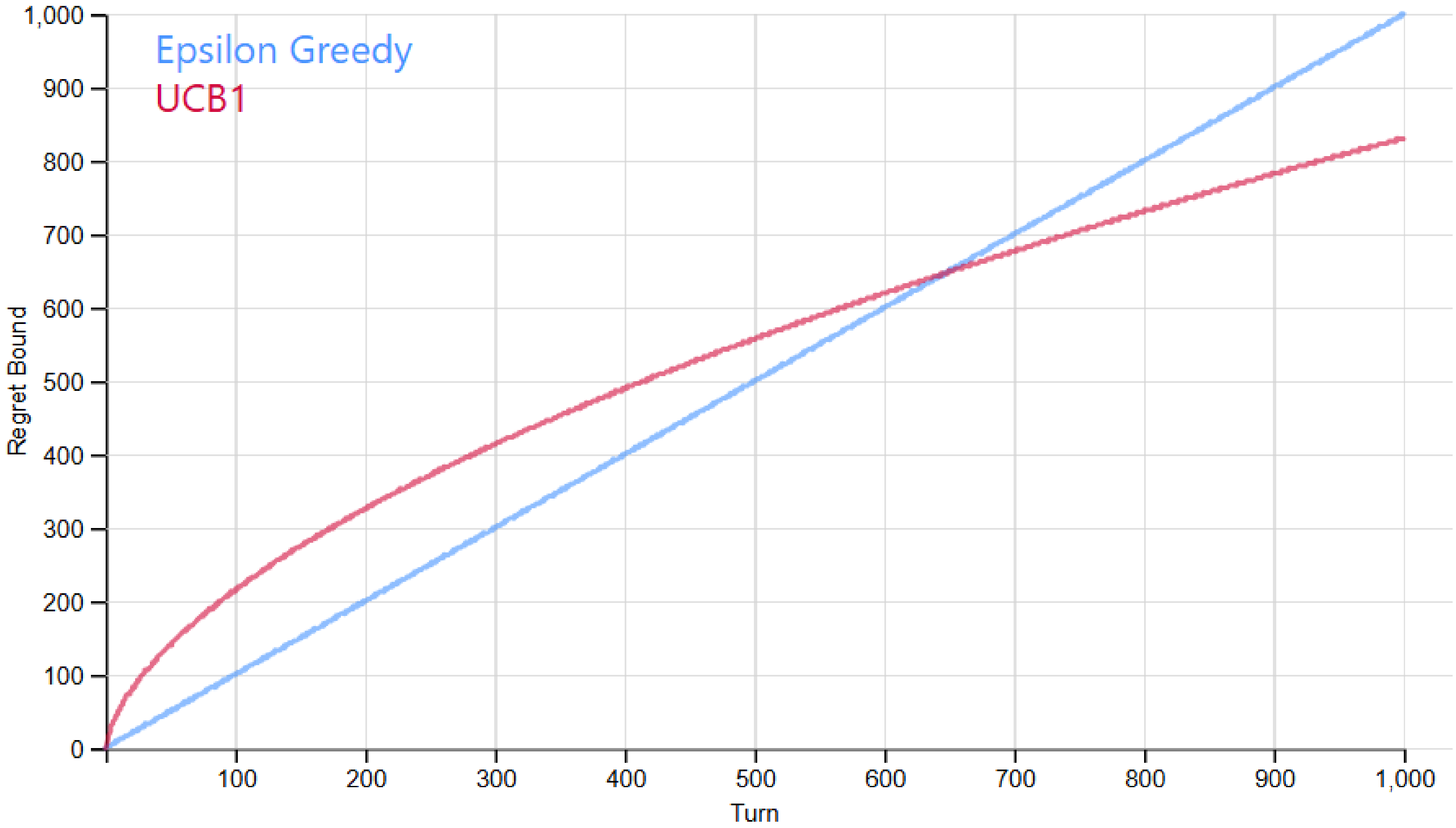
Regret Bound vs. Turn



k (number of arms): 2 T (number of steps): 1000

https://cse442-17f.github.io/LinUCB/

Regret Bound vs. Turn



k (number of arms): 100 T (number of steps): 1000

https://cse442-17f.github.io/LinUCB/

Other Bandit Topics

- Thompson Sampling
- Best Arm Identification
- Adversarial Bandits
- Contextual Bandits
  - State information, s_t
  - Reward depends on state, and action
- Linear Bandits
  - Type of contextual bandit
  - Reward is a linear combination of state features.

information about a learner/a person on the web

$r(s, a)$

~~RL $s_0\ a_0\ s_1\ a_1 \ldots$~~ r_0 r_1

$s_0, a_0 \to r_0$

$s_1, a_1 \to r_1$

known

$$r(s,a) = \theta^T \phi(s,a) = \sum_{i=1}^{n} \theta_i \boxed{\phi_i(s,a)}$$