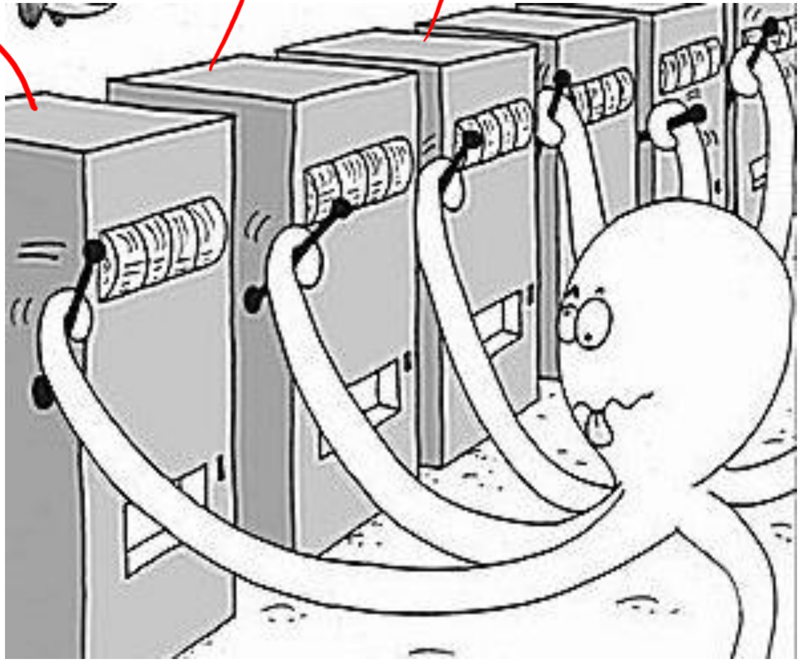
Multi-Armed Bandits

Daniel Brown

r -2 0.5 +1

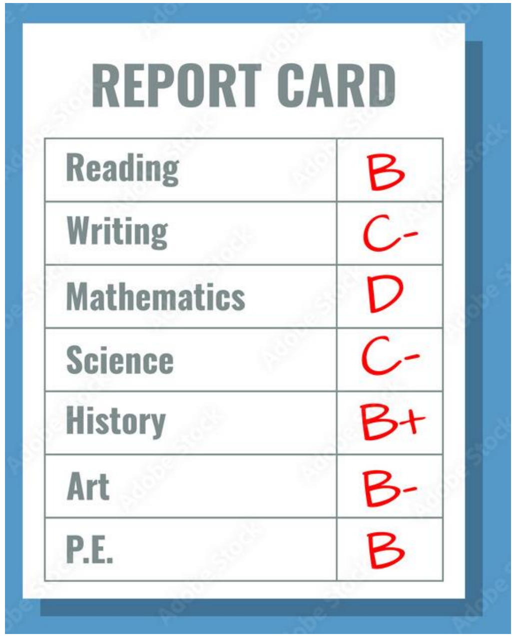


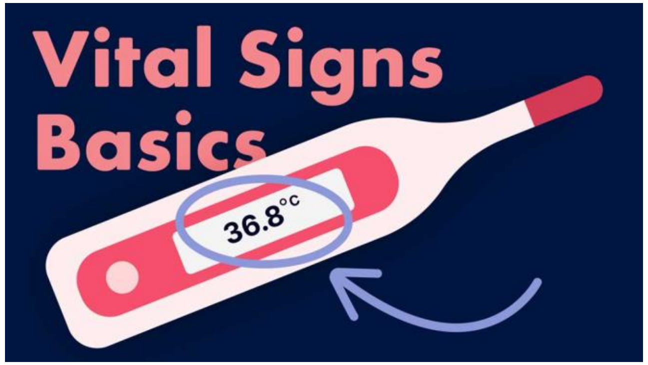
Evaluative feedback

Not supervised learning
– We're not given labels

+1

-1





MAB – reward
– action space

Applications

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

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

Problem formalism

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

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

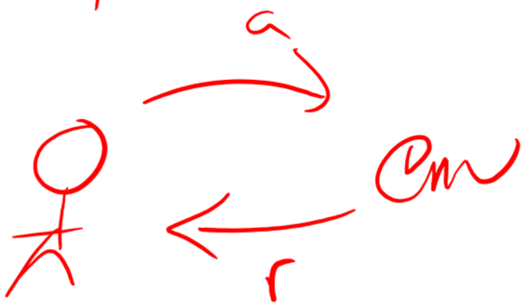
$r(a_3) = N(10,100)$

Actions

- Arms $\mathcal{A} = \{a_1, \dots, a_k\}$
  - 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

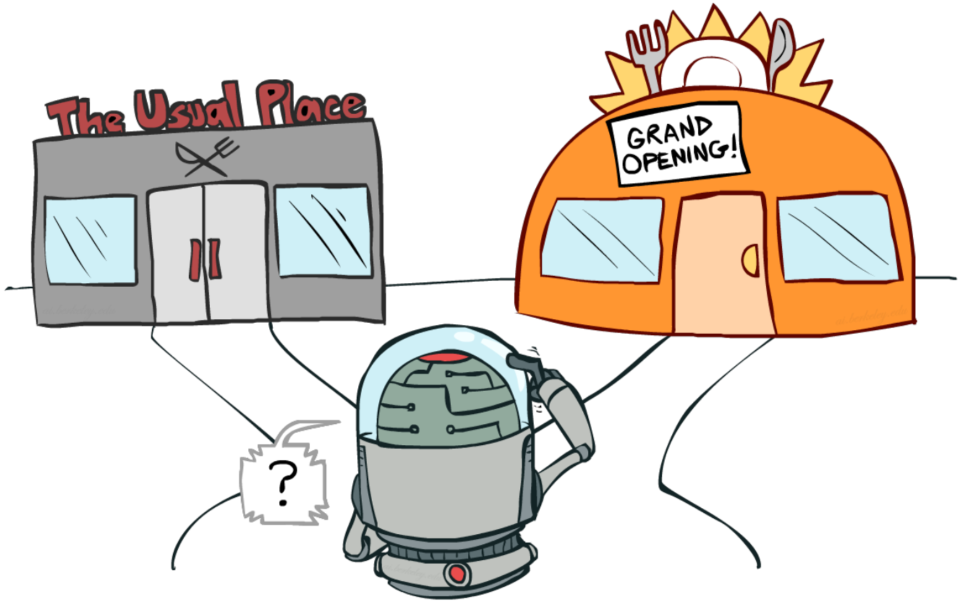
Greedy

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

	a_1	a_2	a_3
$t=1$	0	$t=2$ 0	$t=3$ 1
$t=4$			1
	$\bar{\mu}_1 = 0$	$\bar{\mu}_2 = 0$	$\bar{\mu}_3 = 0$

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

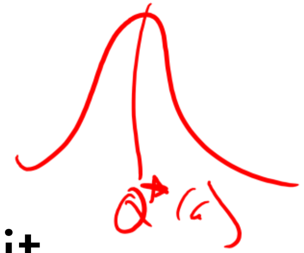
else

take greedy action

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

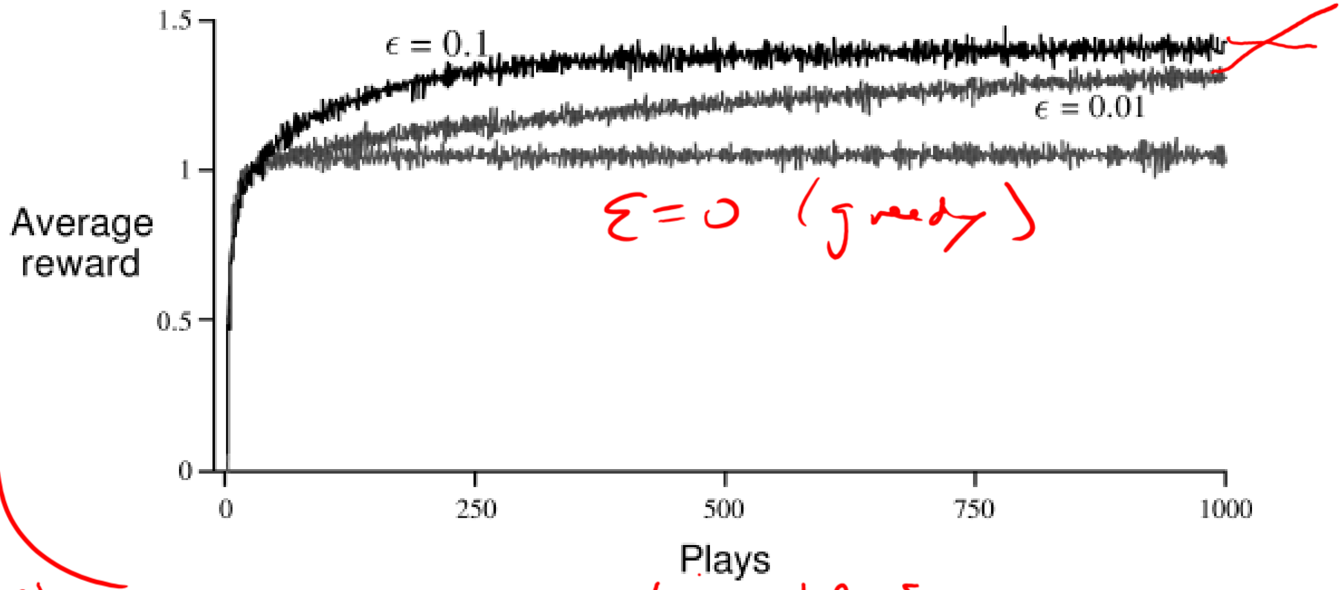
Sutton/Barto figure

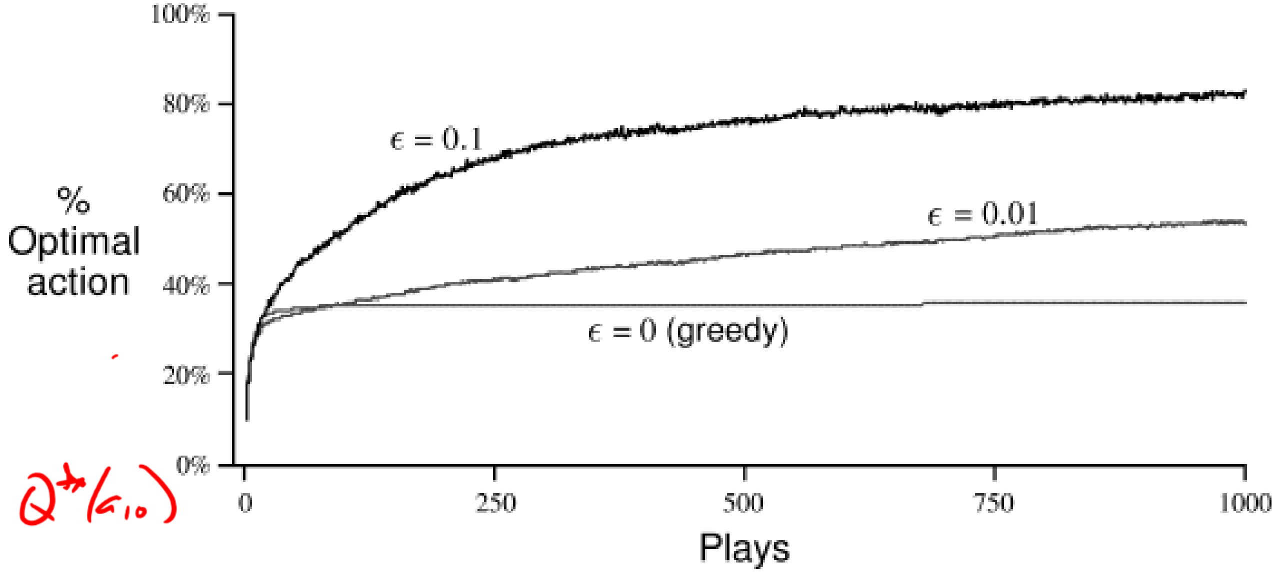
- 10 arms
- Each arm has stochastic reward
  $$r \sim N(Q^*(a), 1)$$
- Averaged over 2000 bandit problems where each problem starts with $Q^*(a) \sim N(0,1)$ for all a



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

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

- Let X be a random variable in the range [0,1] and $x_1, x_2, \dots, x_n$ 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 $P(\bar{X} \geq \mathbb{E}[X] + c) \leq e^{-2nc^2}$

Some fun math

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

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

What is c in terms of δ?

More math

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

What is c?

UCB1 (UCB = Upper Confidence Bound)

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 $\arg\max \bar{X}_i + c(i,t)$
  - Where c$(i,t) = \sqrt{\frac{log(t)}{n_i}}$

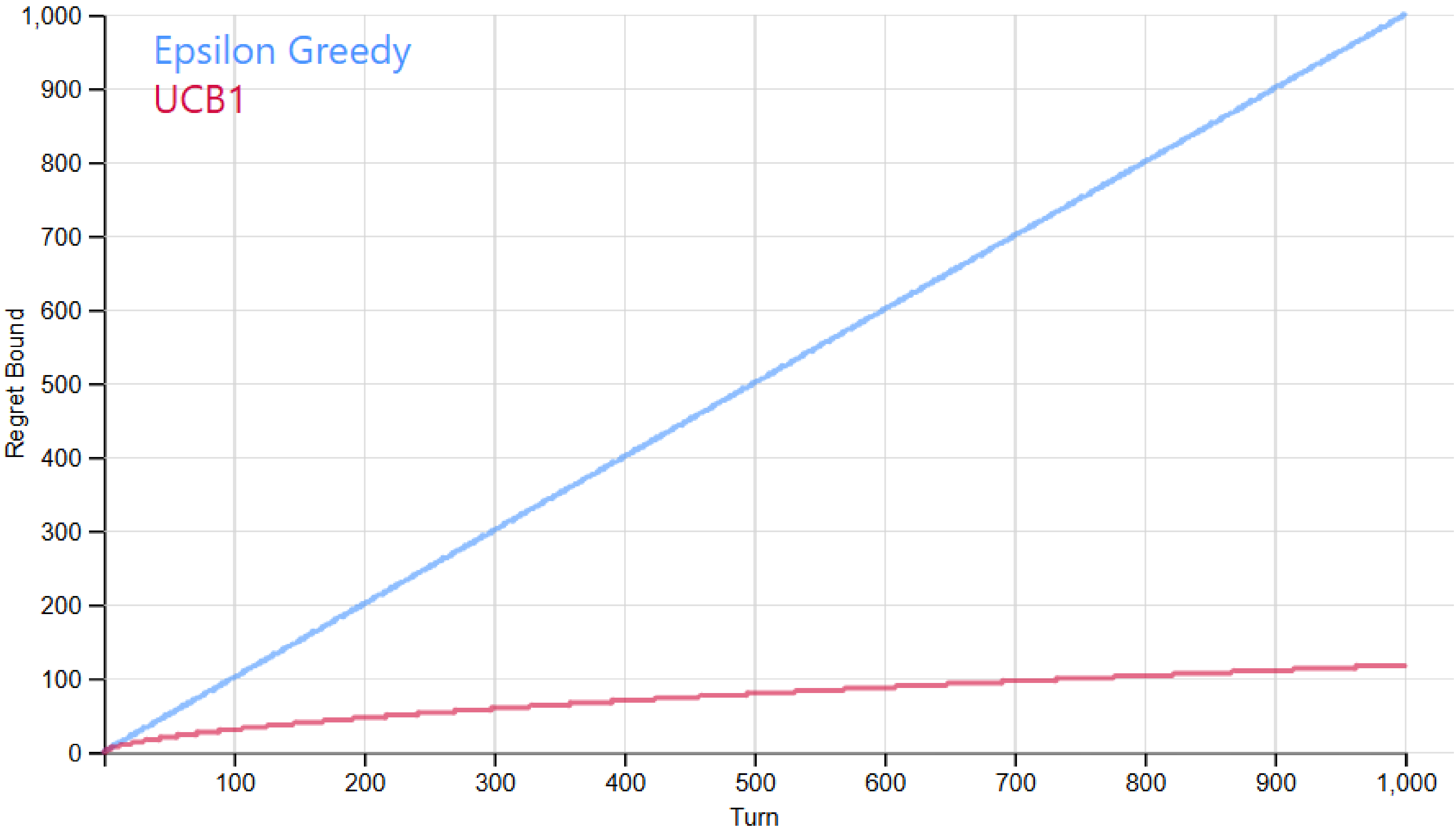
Regret

- 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
  - $O(\sqrt{kTlog(T)})$

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

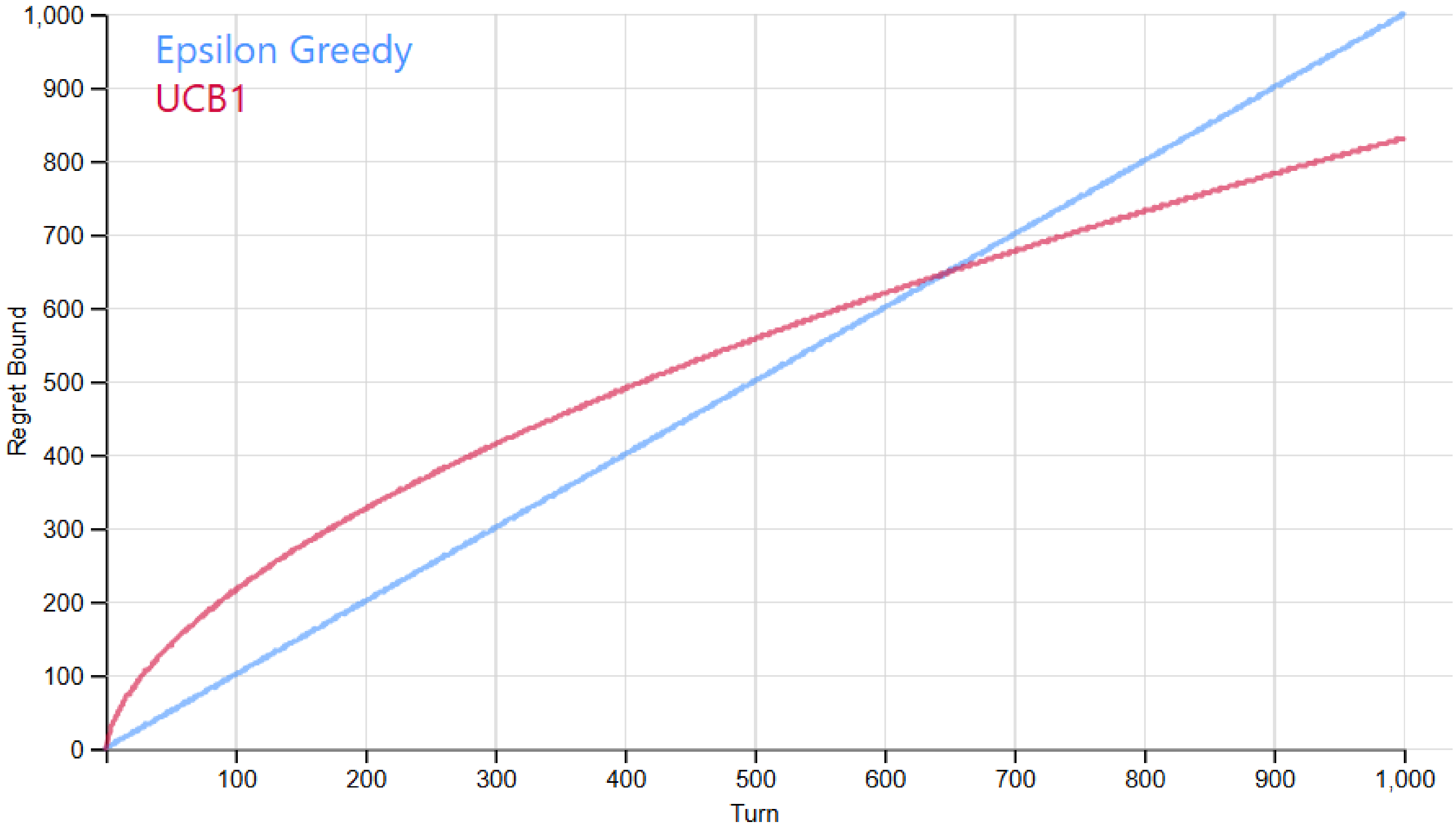
Regret Bound vs. Turn



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.