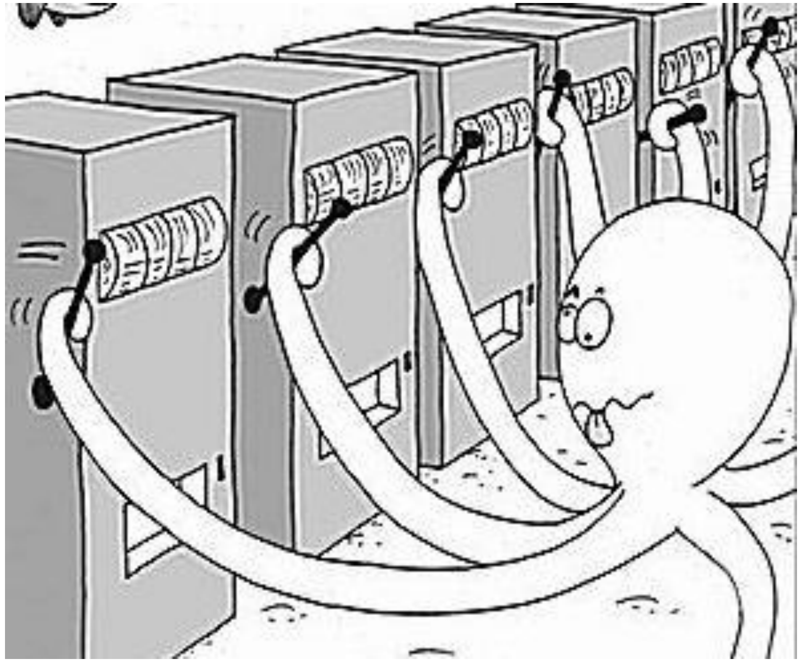
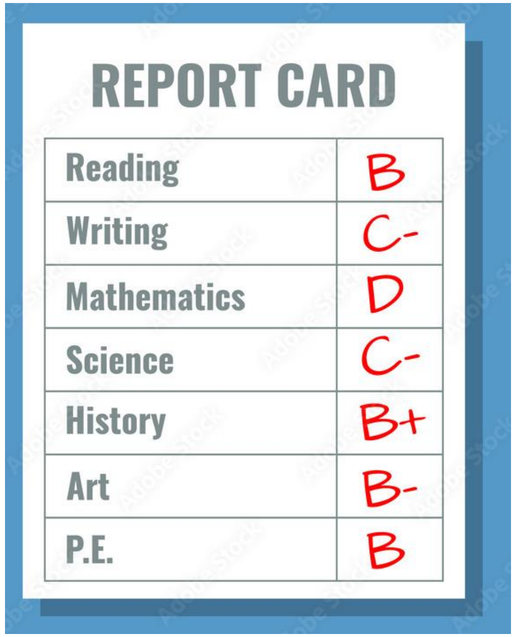
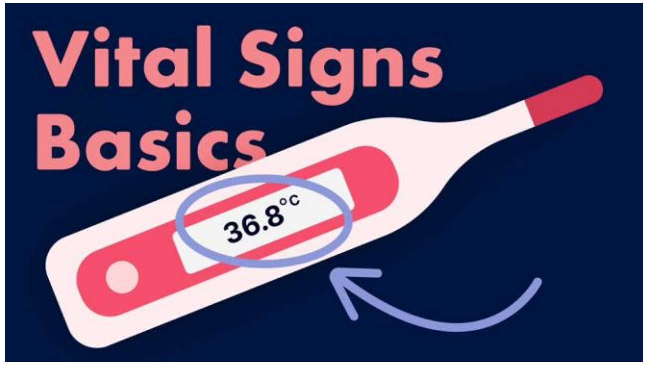
Multi-Armed Bandits

Daniel Brown



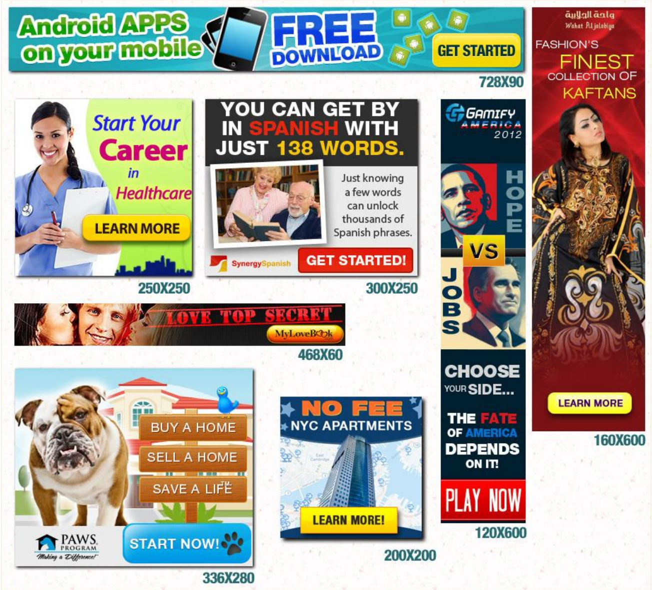
Evaluative feedback





Applications

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



Problem formalism

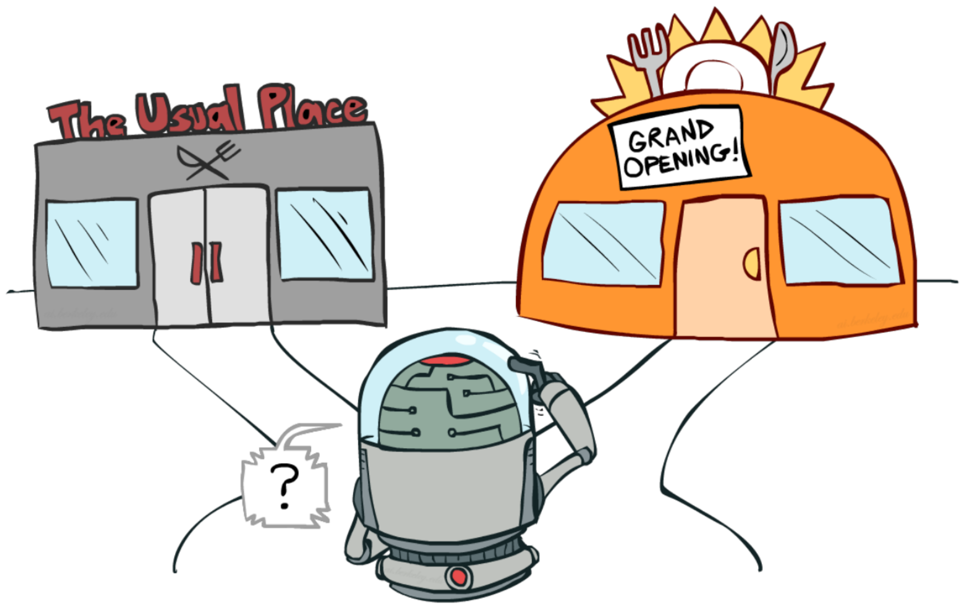
- Arms $\mathcal{A} = \{a_1, \ldots, a_k\}$
  - Each arm is associated with an unknown reward distribution
- Rewards $r_t(a_i)$
- Possible Goals
  - Maximize cumulative reward (Minimize regret)
  - Best arm identification
- Assumptions
  - Independence: Rewards from each arm are independent
  - Stationarity: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

How should we solve this problem?

Random

Greedy

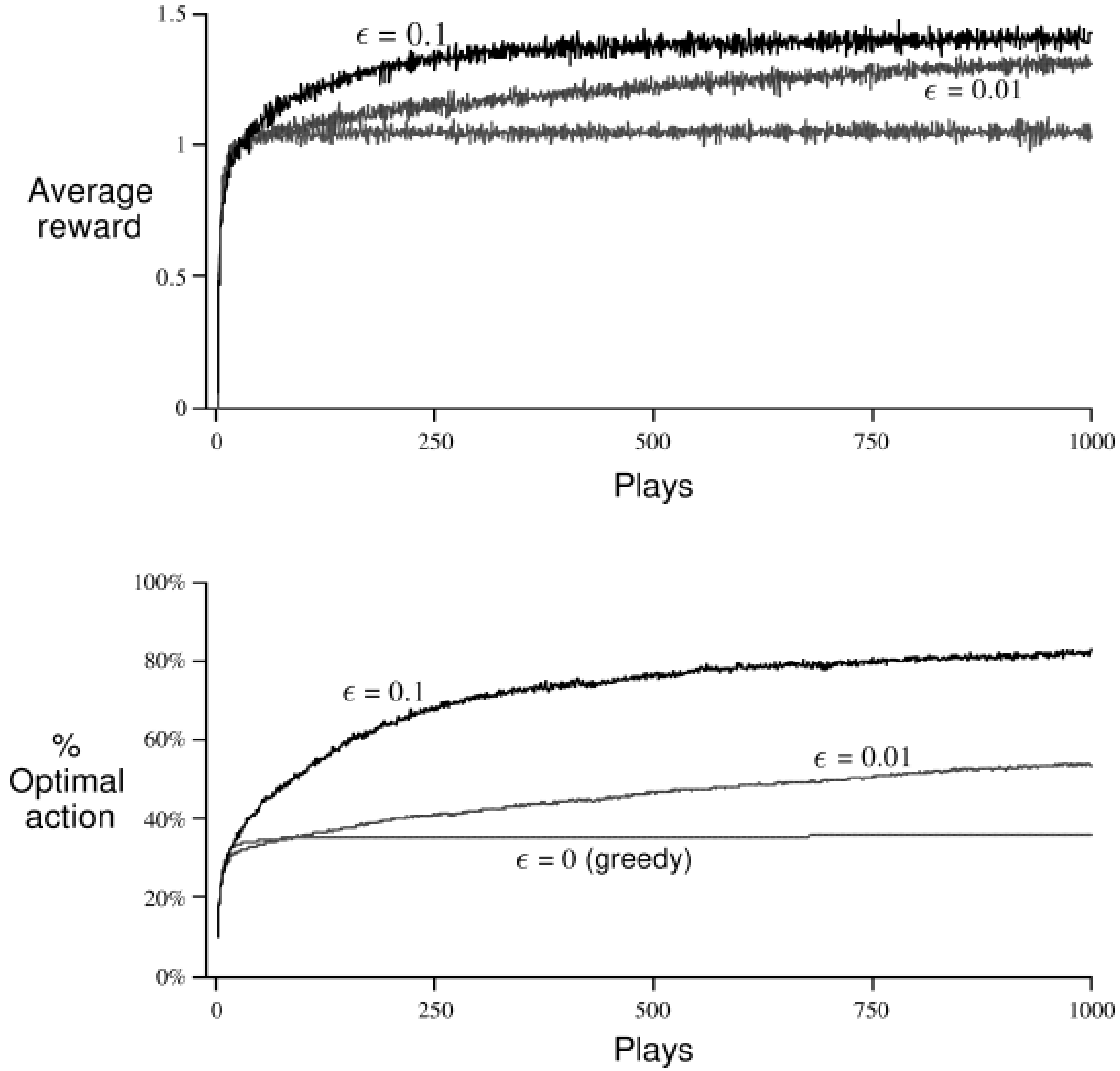
Exploration



ϵ-Greedy

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



Problems?

Boltzmann (Softmax) Exploration

Chernoff-Hoeffding Inequality

- Let $X_1, X_2, \dots, X_n$ be independent random variables in the range [0,1]
- 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?

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

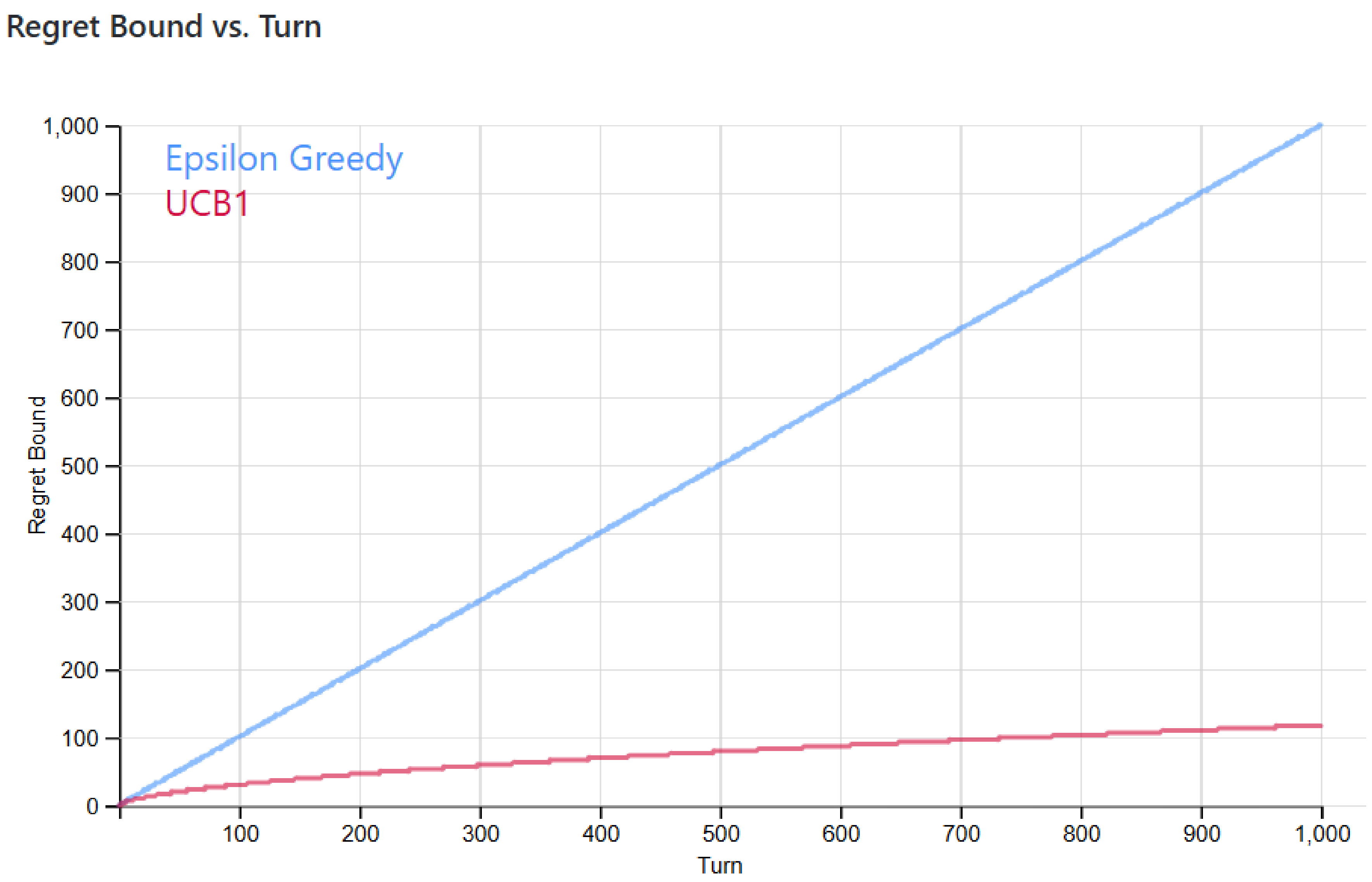
- Play each action once to get initial averages of arm values
- Keep track of counts for each arm n_i
- At each step t select $\arg\max \bar{X}_i + c(i, t)$
  - Where c$(i, t) = \sqrt{\frac{log(t)}{n_i}}$

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

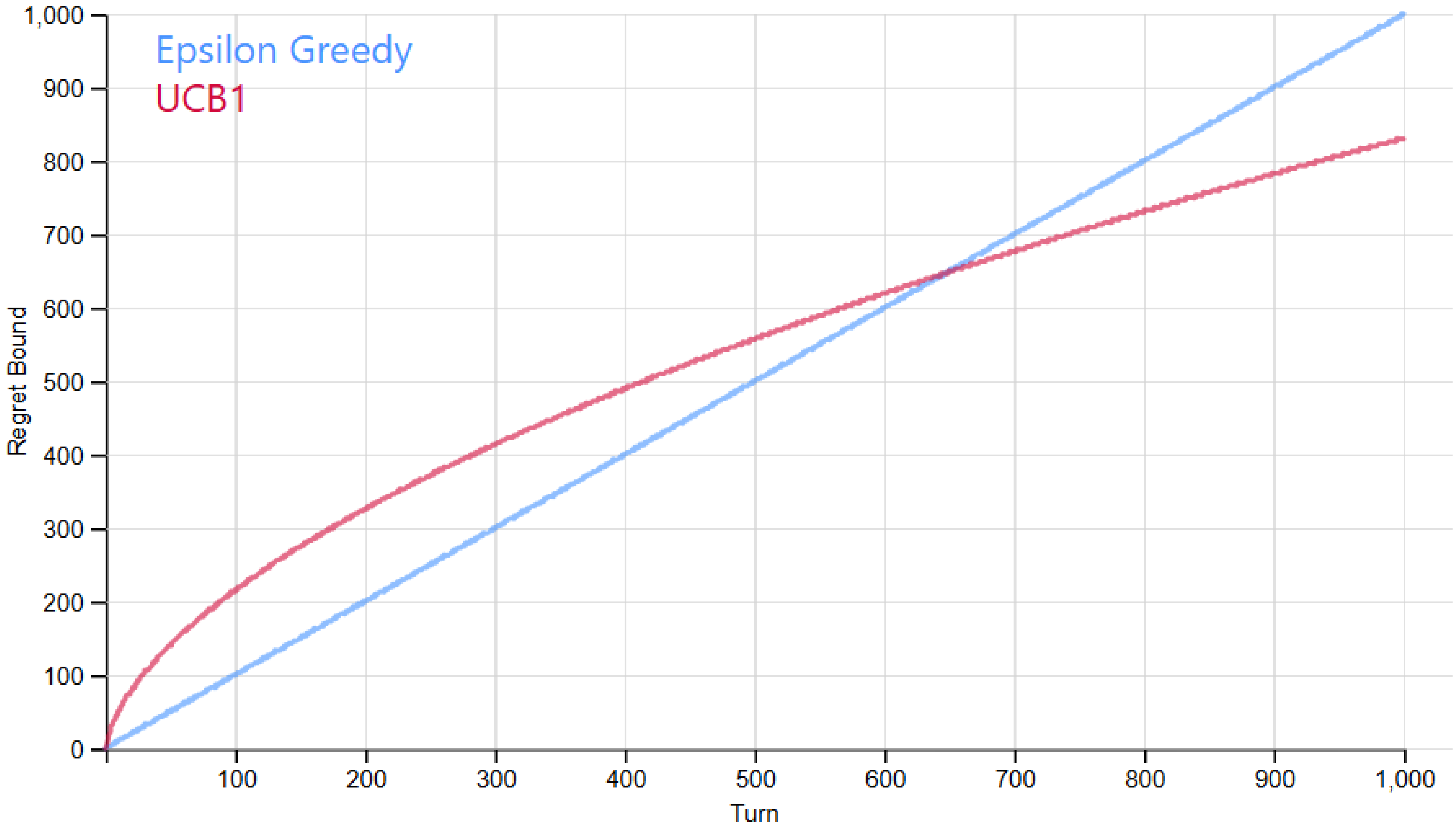
- A **No-Regret** algorithm is such that Regret(T)/T $\rightarrow 0$ as $T \rightarrow \infty$
  - Average regret goes to zero



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.