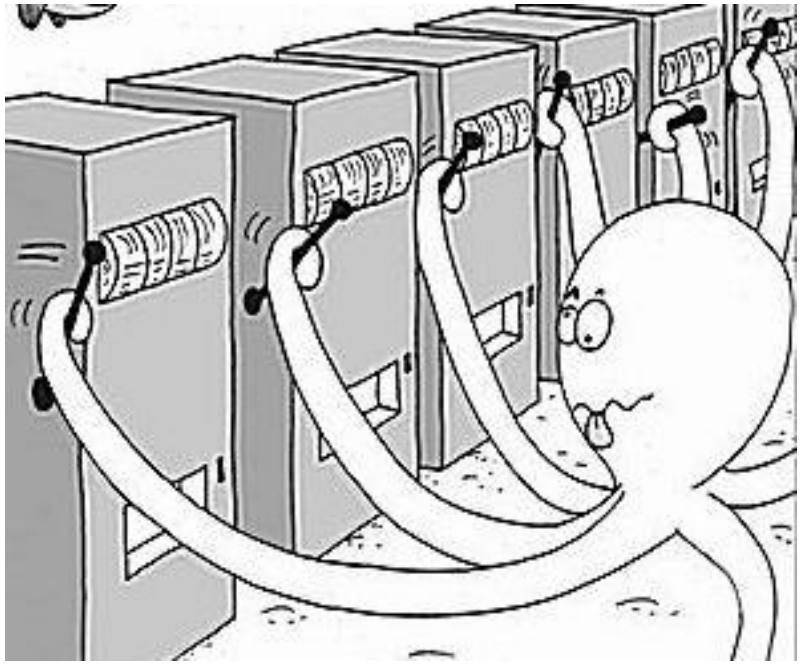
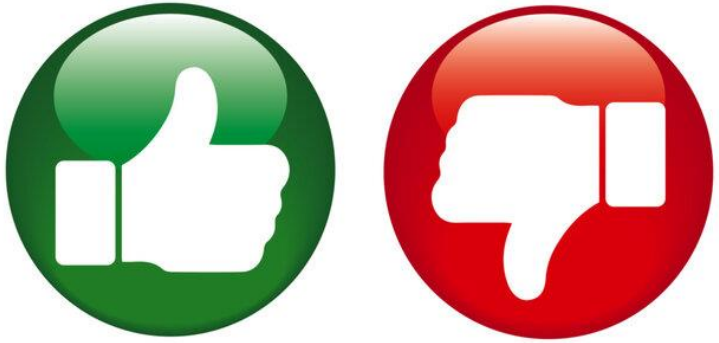
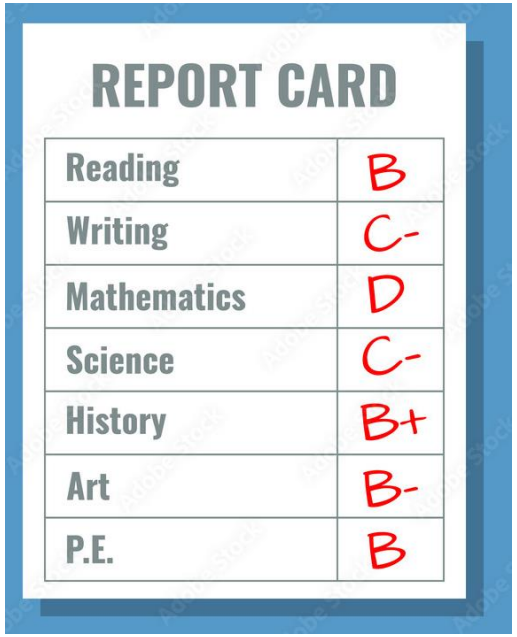
Multi-Armed Bandits

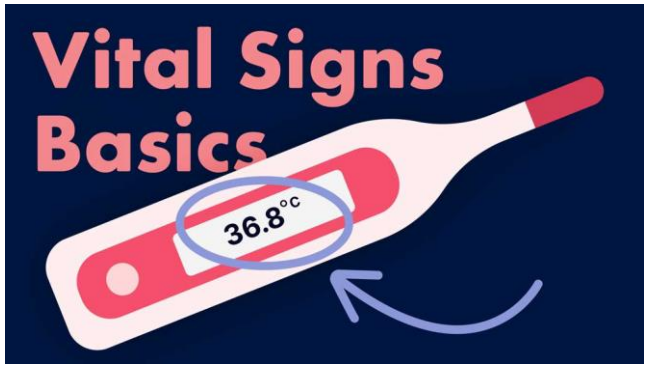
Daniel Brown



Evaluative feedback

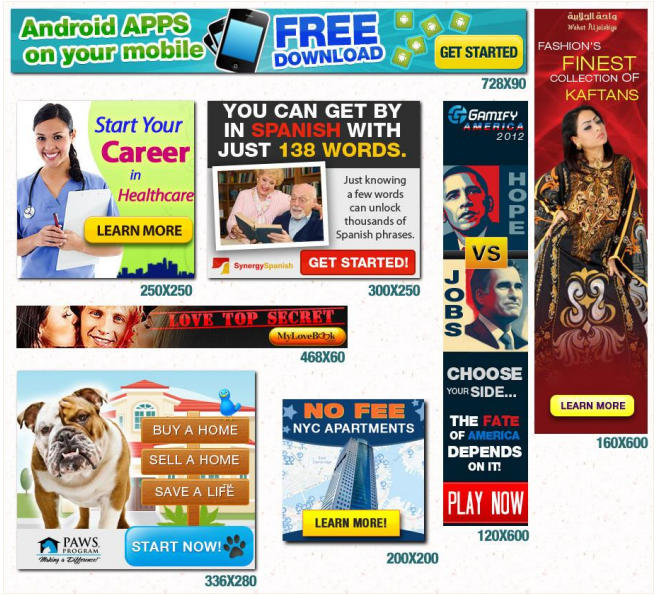






Applications

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



Problem formalism

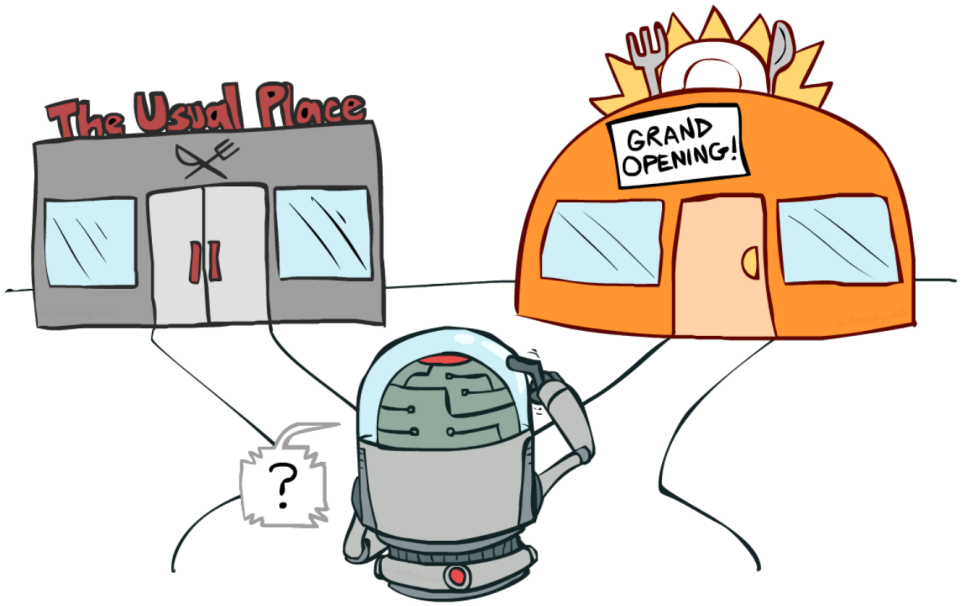
- Arms $\mathcal{A} = \{a_1, \ldots, a_k\}$
  - Each arm is associated with an unknown reward distribution
- Rewards $r_t(a_i)$
- Possible Goals
  - Maximize cumulative reward (Minimize regret)
  - Best arm identification
- Standard Assumptions
  - Independence: Rewards from each arm are independent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

How should we solve this problem?

Random

Greedy

Exploration



ϵ-Greedy

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, $q_*(a)$, $a = 1, \ldots, 10$,

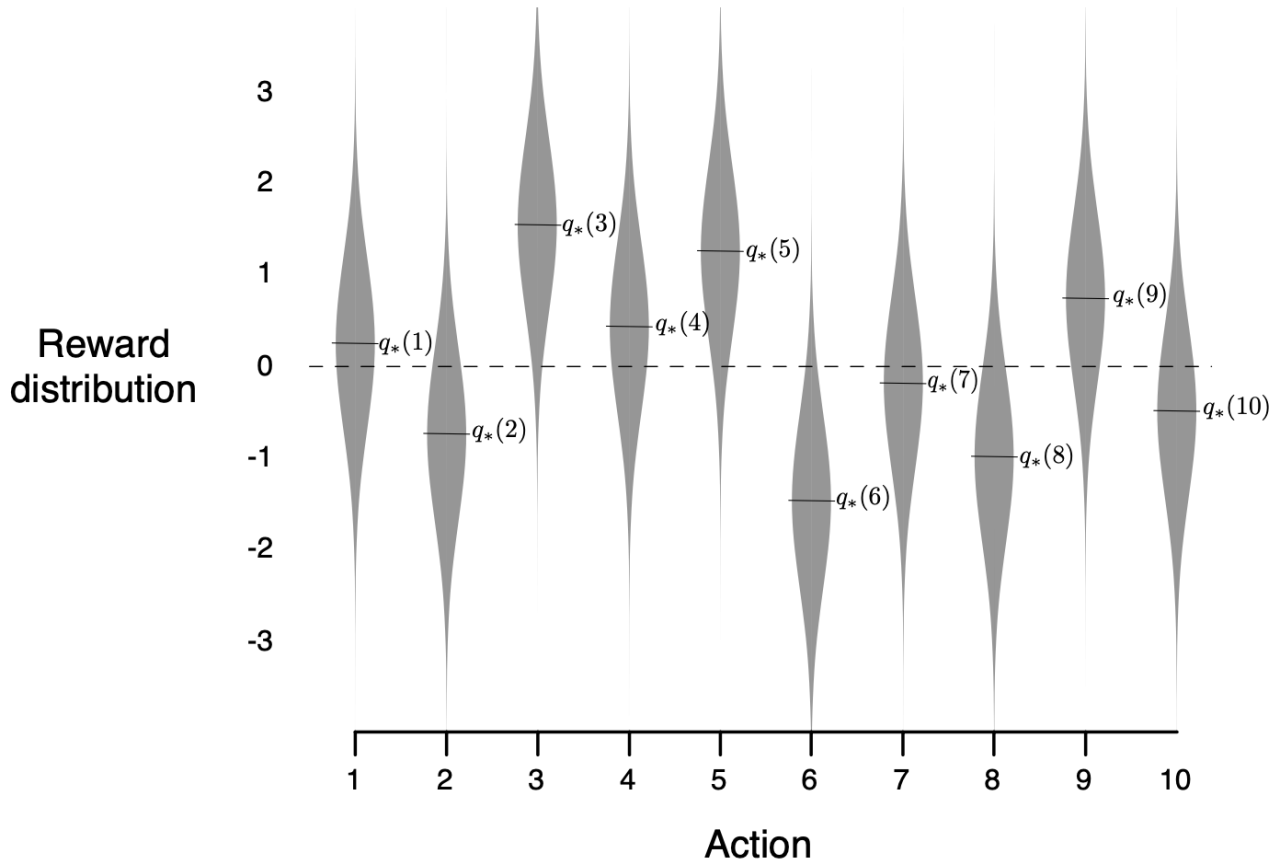


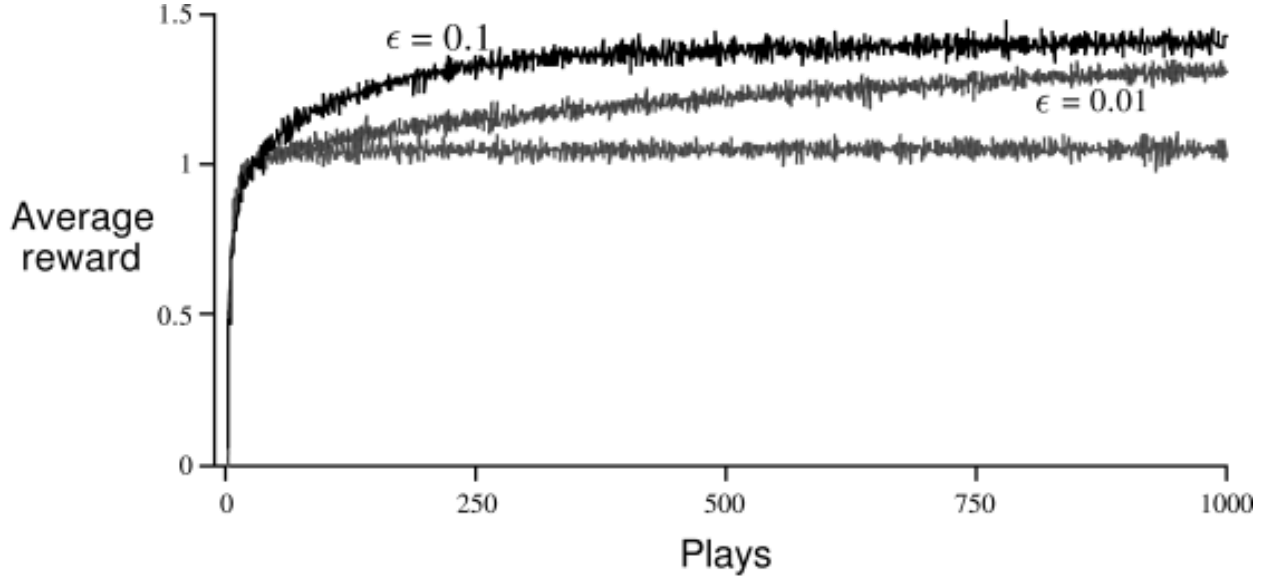
Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution,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

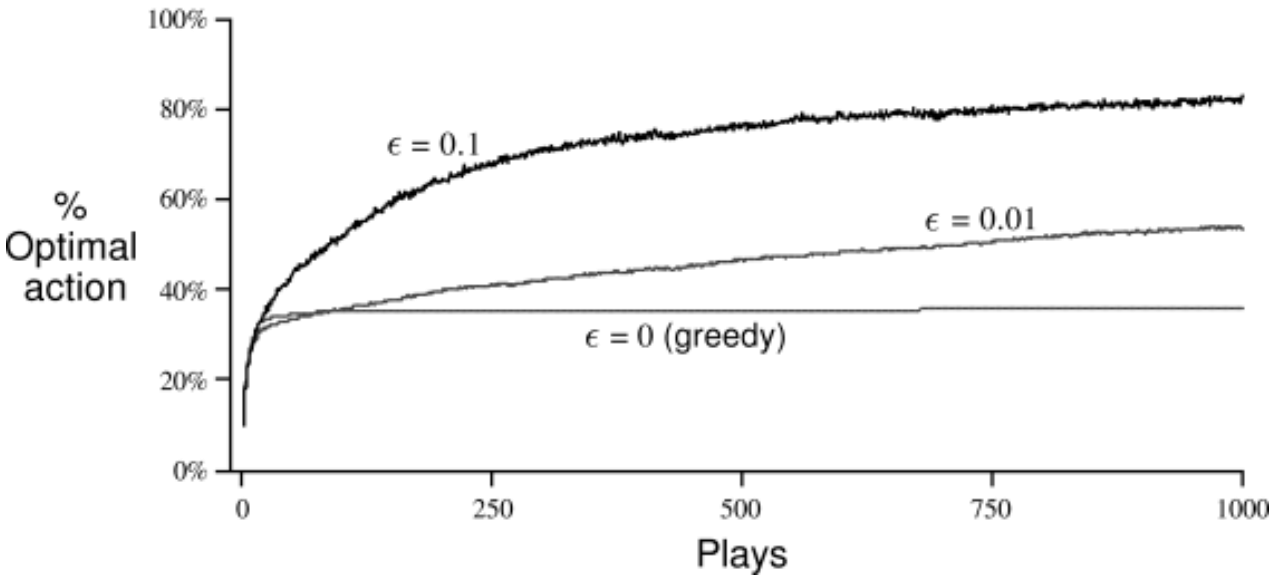
Sutton/Barto figure

- 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





Problems?

Boltzmann (Softmax) Exploration

Chernoff-Hoeffding Inequality

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

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 $\delta = t^{-4}$

What is c?

UCB1 (UCB = Upper Confidence Bound)

Key Idea: Optimism in the face of uncertainty

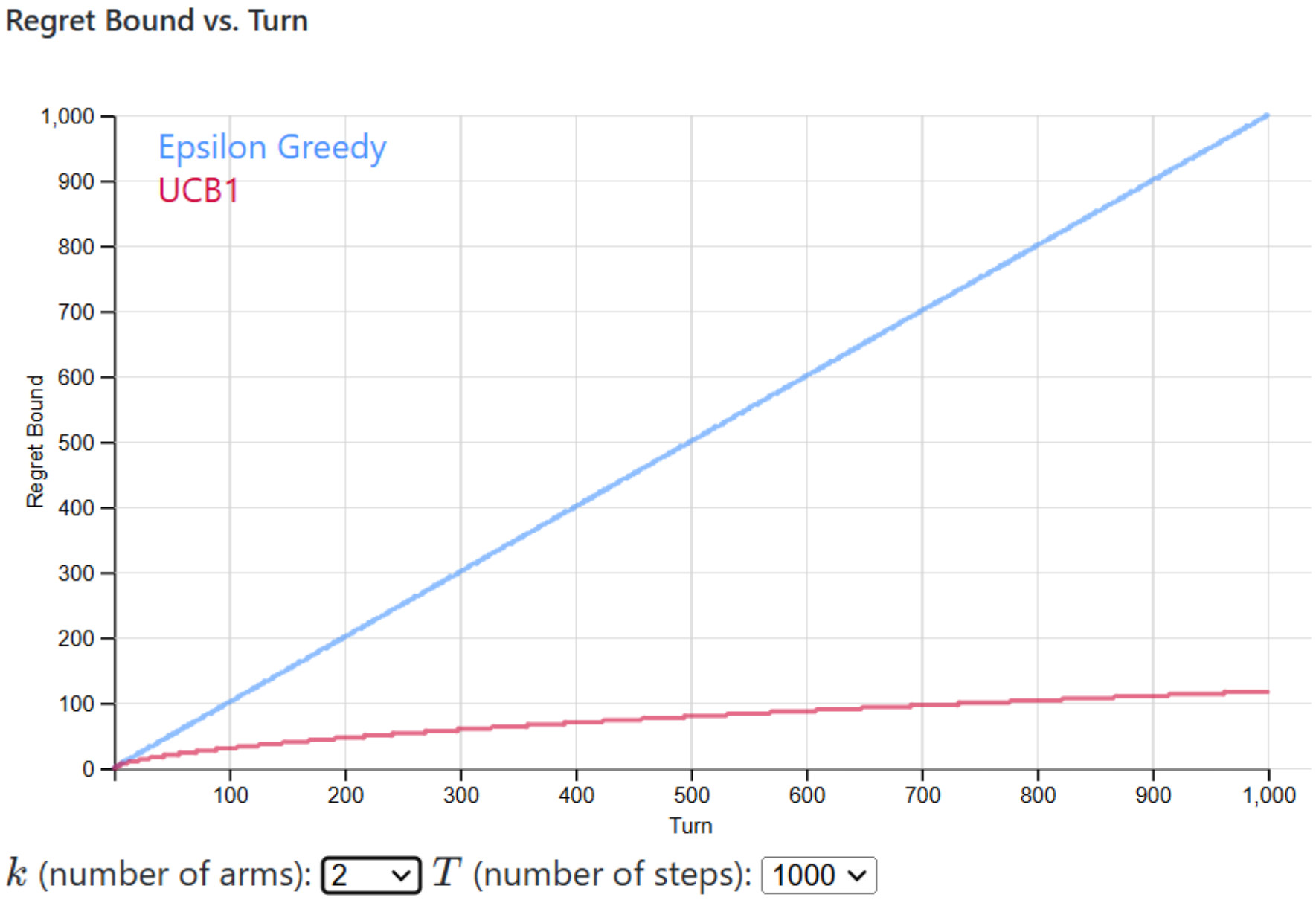
- Play each action once to get initial averages of arm values
- Keep track of counts of pulls for each arm n_i
- At each step t, select $\arg\max \bar{X}_i + c(i,t)$
  - Where $c(i,t) = \sqrt{\frac{2 \cdot log(t)}{n_i}}$

Regret

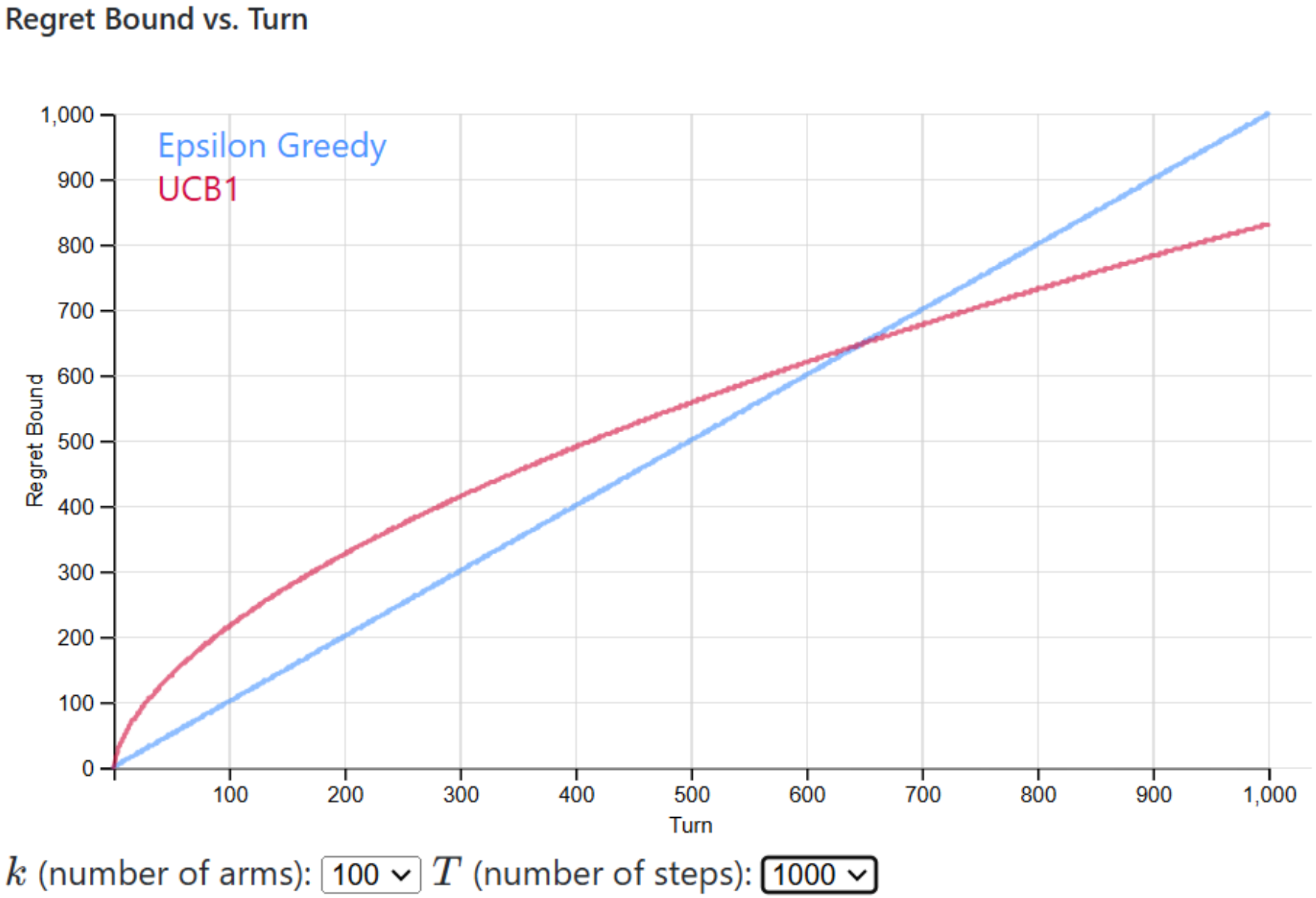
- Define μ^* as the maximum expected payoff over all k arms
- Regret(T) = $T\mu^* - \sum_{t=1}^{T} r_t$

- 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



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



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

Notes

- The version we derived is for rewards in range [0,1]
- What happens if rewards are in range [a,b]?
  - Just scale the upper confidence value

$$\mathrm{c}(i,t) = (b-a)\sqrt{\frac{2 \cdot log(t)}{n_i}}$$

In practice the scaling term is often just treated as a hyperparameter that controls exploration vs. exploitation.

$$\mathrm{c}(i,t) = \alpha\sqrt{\frac{log(t)}{n_i}}$$

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.