Bandits

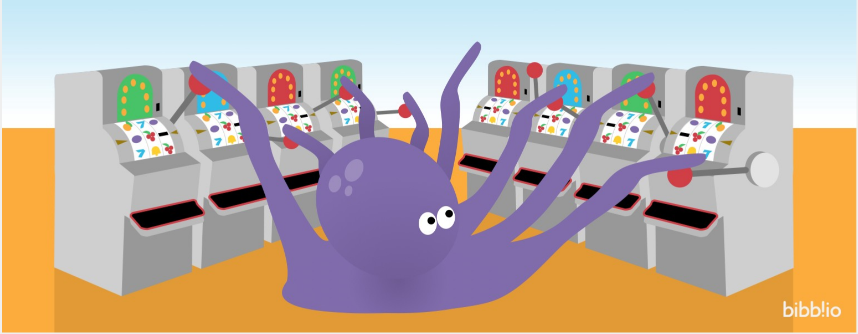
Weizi Li

Department of Computer Science
University of Memphis

- Introduction
- Action-value Method
- Incremental Method

Introduction

- Problem setting: repeatedly choose among k different actions. After each action you receive a numerical reward. Actions have no further influence.
- Goal: maximize the expected total reward over some time steps (e.g., 1000 action selections).
- Example with four arms:
  - Machine 1 (50%)
  - Machine 2 (70%)
  - Machine 3 (35%)
  - Machine 4 (45%)


bibb!io

- When estimating action values, at any time step, there is always one "optimal" action.
- Exploitation: acting greedily to the "optimal" action (short-term benefits).
- Exploration: choosing new actions (potential long-term benefits).
- RL requires a balance between the two.

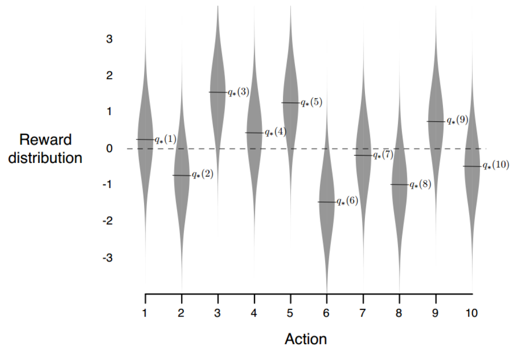
Factors of Bandits

- Current estimates of the action values
- System uncertainties (e.g., stationary vs non-stationary)
- Number of available steps
- Easy to solve if we have the following:
  - actual action values
  - no system uncertainty
  - infinite number of steps

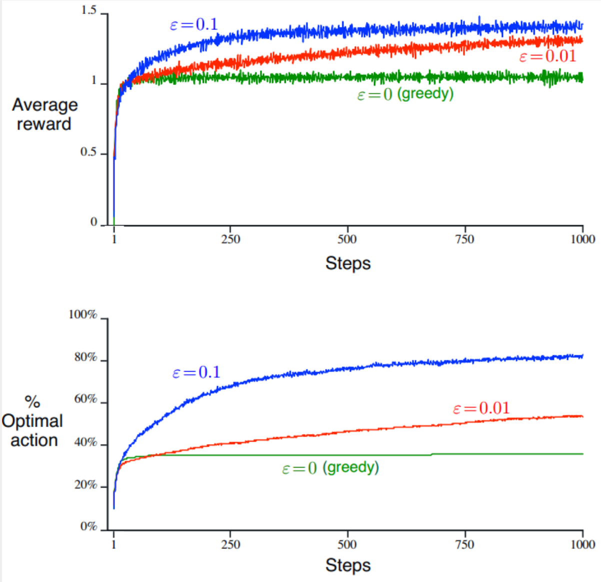
Action-value Method

- Estimating action value by averaging the received rewards.
- Choosing action according to either greedy or ϵ-greedy strategies. The latter approach can work surprisingly well, but the performance is task-dependent.
- Solution is approximated, since we do not have infinite number of time steps.

- First sample $\mathcal{N}(0,1)$ to get the actual expected action values $\{Q(a)\}_{a=1}^{10}$
- Execute the action $a \in [\![1, 10]\!]$ to receive $\mathcal{N}(Q(a), 1)$ reward
- Run for 1000 steps



Performance over 2000 Tasks



Incremental Method

- Estimate action value incrementally, instead of computing the average in the end:

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\
&= Q_n + \frac{1}{n}\left[R_n - Q_n\right]
\end{aligned}
$$

- Choosing action according to either greedy or ϵ-greedy strategies.

- $Q(a)$ changes over time. So, it's better to put more weight to recent rewards than to long-past rewards.
- We can use a constant (or dynamic) step-size parameter (α).

$$Q_{n+1} = Q_n + \alpha \left[R_n - Q_n \right]$$