CS774 Reinforcement Learning

Lecture #2: Evaluative Feedback

Kee-Eung Kim
KAIST EECS Department
Computer Science Division

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Introduction

☐ Evaluating actions vs. instructing by giving correct actions

- Pure evaluative feedback depends totally on the action taken (how good the action taken is, no indication of best or worst).
  Pure instructive feedback depends not at all on the action taken (tells correct action to take).
- Supervised learning is instructive; optimization is evaluative

☐ Associative vs. Nonassociative:

- Associative: inputs mapped to outputs; learn the best output for each input
- Nonassociative: "learn" (find) one best output

☐ *n*-armed bandit problem (at least how we treat it) is:

- Nonassociative
- Evaluative feedback

The *n*-Armed Bandit Problem

- Select repeatedly from one of *n* actions;
  each selection is called a play
- After each play a_t,
  you get a reward r_t, where
  $E[r_t|a_t] = Q^*(a_t)$
  - These are values of actions unknown to the agent
  - Distribution of r_t only depends on a_t
- Objective
  - Maximize the total reward over long term, e.g., 1000 plays
- Exploration vs. Exploitation
  - To solve the *n*-armed bandit problem, agent must *explore* a variety of actions and *exploit* the best of them

The Exploration/Exploitation Dilemma

☐ Suppose you form estimates

$$Q_t(a) \approx Q^*(a) \quad \text{action-value estimates}$$

- $a_t^* = \text{argmax}_a \, Q_t(a)$
- $a_t = a_t^*$ then exploitation
- $a_t \neq a_t^*$ then exploration

- The agent can't exploit all the time; the agent can't explore all the time

- You can never stop exploring; but you should always reduce exploring. Maybe.

Action-Value Methods

- Methods that adapt action-value estimates and nothing else, e.g.: suppose by the t-th play, action a had been chosen k_a times, producing rewards $r_1, \ldots, r_{k_a}$ then

$$Q_t(a) = \frac{r_1 + \cdots + r_{k_a}}{k_a} \qquad \text{sample average}$$

- Convergence: $\lim_{k_a \to \infty} Q_t(a) = Q^*(a)$

ε-Greedy Action Selection

☐ Greedy action selection

- $a_t^* = \text{argmax}_a Q_t(a)$
- $a_t = a_t^*$

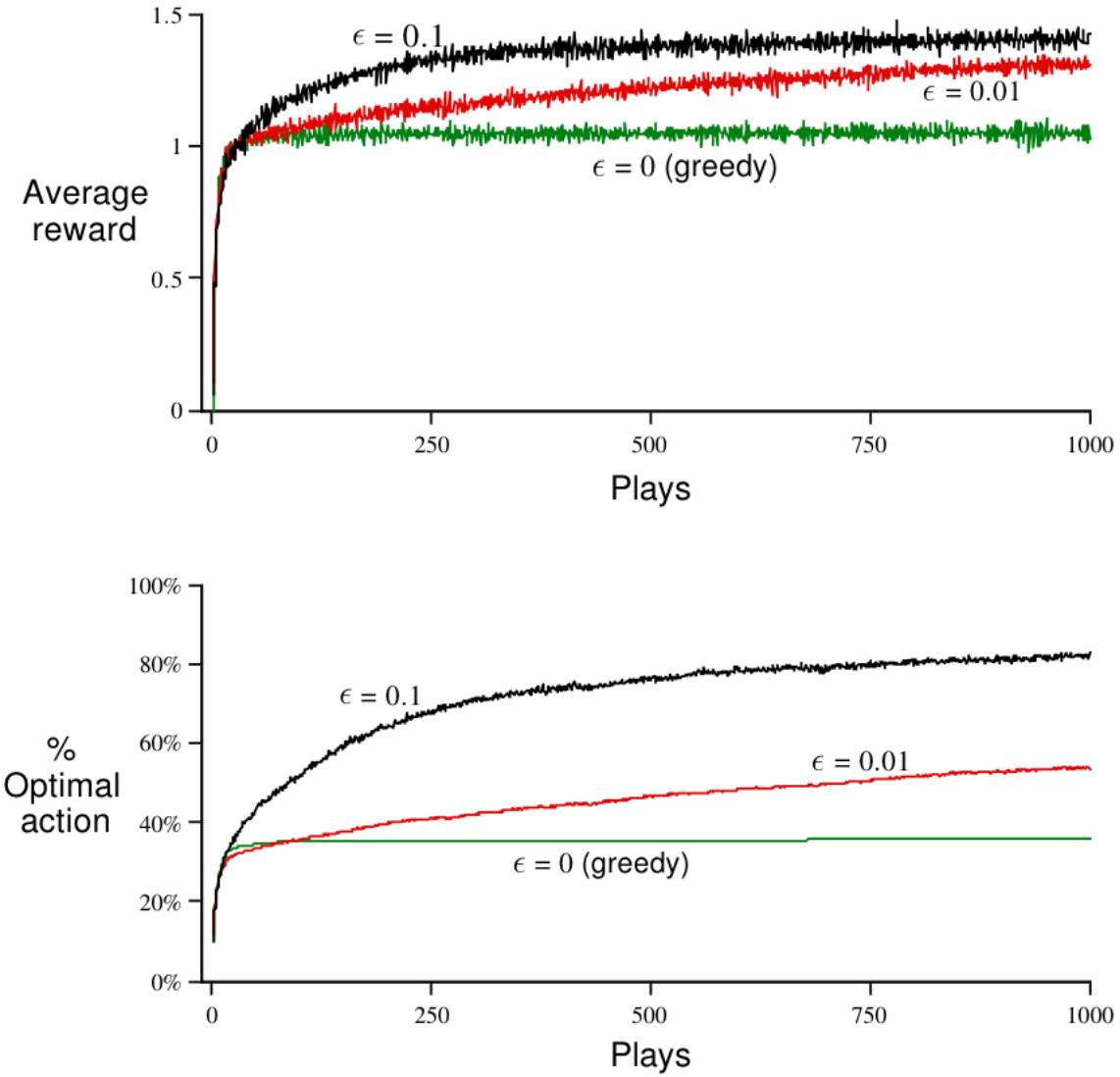
☐ ε-Greedy Action Selection

- $$a_t = \begin{cases} a_t^* \text{ with probability } 1-\epsilon \\ \text{random action with probability } \epsilon \end{cases}$$

10-Armed Bandit

- n = 10 possible actions
- Each $Q^*(a)$ is chosen randomly from a normal dist: $N(0, 1)$
- Each r_t is also normal: $N(Q^*(a), 1)$
- 1000 plays
- repeat the whole thing 2000 times and average the results

ε-Greedy Method for 10-Armed Bandit



Softmax Action Selection

- Softmax action selection methods grade action probabilities by estimated values
- The most common softmax uses a Gibbs (also called Boltzmann) distribution
  - Choose action a on the t-th play with probability

$$\frac{e^{Q_t(a)/\tau}}{\sum_{a'} e^{Q_t(a')/\tau}}$$

where τ is a positive parameter called temparature

Incremental Implementation

☐ Recall the sample average estimation method:

$$Q_t(a) = \frac{r_1 + \cdots + r_{k_a}}{k_a}$$

- We need to store k_a rewards
- Can we do this incrementally (without storing all the rewards)?

☐ We could keep a running sum and count:

- Count: increment k_a whenever action a is selected
- Running sum:

$$Q_{k_a+1}(a) = Q_{k_a}(a) + \frac{1}{k_a + 1}[r_{k_a+1} - Q_{k_a}]$$

☐ Common form for update rules

$$NewEstimate \leftarrow OldEstimate + StepSize[Target - OldEstimate]$$

Tracking a Nonstationary Problem

☐ What if the bandit changes over time?

- We encounter the nonstationarity more often than you can imagine

☐ Give more weights on the recent rewards

- (Sample average = equal weights)
- How?
  - Update with constant step-size parameter α:

$$Q_{k+1} = Q_k + \alpha[r_{k+1} - Q_k]$$

- Why?

$$\begin{aligned} Q_k &= Q_{k-1} + \alpha[r_k - Q_{k-1}] \\ &= \alpha r_k + (1-\alpha)Q_{k-1} \\ &= \alpha r_k + (1-\alpha)\alpha r_{k-1} + (1-\alpha)^2 Q_{k-2} \\ &= \alpha r_k + (1-\alpha)\alpha r_{k-1} + (1-\alpha)^2 \alpha r_{k-2} + \cdots \\ &\quad + (1-\alpha)^{k-1}\alpha r_1 + (1-\alpha)^k Q_0 \\ &= (1-\alpha)^k Q_0 + \textstyle\sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} r_i \end{aligned}$$

Optimistic Initial Values

- ☐ Methods so far dependent to $Q_0(a)$, i.e. biased
- ☐ Suppose we initialize them to +5 (optimistic)
  - Optimism encourages action-value methods to explore!
  - Why?
    - We get "disappointed" by the actual reward received, and switch to other action. Switch = exploration