Reinforcement learning in neuroscience - tutorial

Rita Almeida

PhD course 3045, VT 2018

Karolinska Institutet

n-armed bandit problem

- Repeated plays - action selections.
- Each play *n* possible actions.
- After an action a reward is received.
- Each action is associated with a stationary probabilistic reward.
- Aim is to achieve maximal reward over a number of plays (for example 1 000).

Exploitation and exploration

Greedy action: action with greatest estimated value.

Exploitation: selecting a greedy action - exploiting what is known about the values.

- maximizes expected reward on the next play

Exploration: selecting a non-greedy action - exploring to improve estimates of values

- can lead to greater total reward
- immediate reward is on average lower

Estimating action-values $Q(a)$

- Sample-average method: estimating $Q(a)$ as the average reward previously received when choosing a.

  If at step t action a has been chosen k_a times:

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}$$

- Incremental method: updating the estimated $Q(a)$ after each step.

  If Q_k is the average of the first k rewards

$$Q_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$$

Update rule

$$Q_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$$

Generic update rule:

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

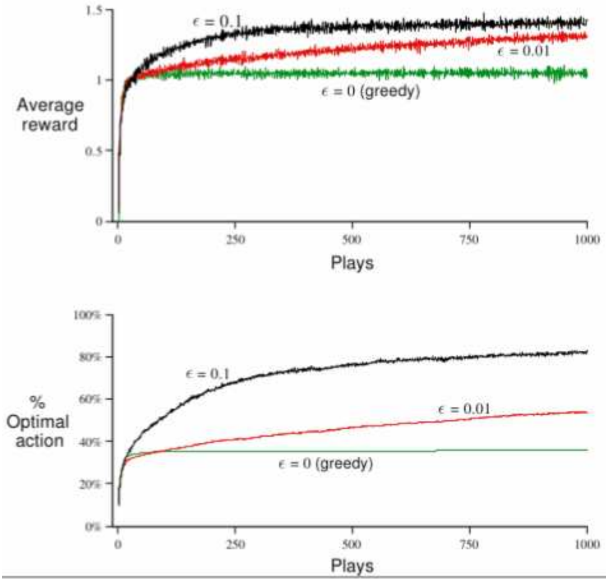
- Similar rules are common in RL.
- $[Target - OldEstimate]$: error of the estimate
- $LearningRate$: size of update

Action selection

- Greedy method: select action a^* with maximal average past reward
- ϵ-greedy method: behave greedily but with small probability ϵ choose a non-greedy action
  - If one plays infinitely ($k_a \to \infty$) selecting the optimal action converges to greater than $1 - \epsilon$

10-armed bandit example

- 10-armed bandit
- for each a reward is chosen from a normal distribution with mean $Q^*(a)$ and variance 1
- 1000 plays
- repeat everything 2000 times and average
- $\epsilon = 0.01$ improves more slowly
- $\epsilon = 0.01$ is better in the long run
- maybe method where ϵ decreases with time



ϵ-greedy method

- As the reward variance increases more exploration is needed to get greater total reward.
- If reward variance is 0 then it is enough to select each action 1 time to estimate the rewards.
- If the reward distribution is non-stationary exploration is needed to get greater total reward.
- ϵ-greedy methods choose all actions with equally probability when exploring.

Softmax action selection

- Idea: probability of choosing an action is proportional to its estimated value

Action a is chosen on play t with probability:

$$\frac{e^{Q_t(a)\beta}}{\sum_{b=1}^{n} e^{Q_t(b)\beta}}$$

- $\beta \geq 0$ is the inverse temperature
- $\beta \to 0$: all actions are chosen with similar probabilities
- large β: greedy action selection

2-armed bandit

Example: Subject chooses between left or right (L or R).

$$P(L)_t = \frac{e^{\beta Q_t(L)}}{e^{\beta Q_t(L)} + e^{\beta Q_t(R)}} \qquad P(R)_t = \frac{e^{\beta Q_t(R)}}{e^{\beta Q_t(L)} + e^{\beta Q_t(R)}}$$

$$P(L)_t = \frac{1}{1 + e^{-\beta(Q_t(L) - Q_t(R))}}$$