Date created: 2025-07-15-Tue

# 2.3 The 10-armed Testbed

---

## Experiment Setup

To test the effectiveness of greedy and $\epsilon$-greedy methods (as well as other methods), we set up the following experiment:

Each experiment:

- has 10 possible actions $a = 1, 2, ..., 10$,
- each $q_*(a) \sim N(0, 1)$,
- when $A_t$ is selected, the actual reward $R_t \sim N(q_*(A_t), 1)$.

This setup is called the 10-armed testbed, where running each experiment for 1000 steps makes up a run. We repeat 2000 runs and average the results.

## Results

Comparing three methods, greedy, $\epsilon$-greedy with $\epsilon = 0.1$, and $\epsilon$-greedy with $\epsilon = 0.01$, we find that:

- the greedy method found the optimal action only 1/3 of the runs,
- the 0.1-greedy method found the optimal action the fastest, while the 0.01-greedy method took longer to find the optimal action,
- however, in the long run the 0.01-greedy method would accumulate more reward than the 0.1-greedy method since when the optimal action is found for both methods, the 0.01-greedy method would select it 99% of the time whereas the 0.1-greedy method would only select it 90% of the time.

## Analysis

The $\epsilon$-greedy methods are advantageous over the greedy method because the rewards have *variance*. It takes several tries to determine with confidence the true value of an action, highlighting the importance of exporation. Suppose the rewards have zero variance, then it takes only one try to determine the true value of each action. In this case, the greedy method would do better than $\epsilon$-greedy methods.

But even in the deterministic case there would be advantages to the $\epsilon$-greedy methods were we to relax some other assumptions. For example, suppose the bandit tasks are **nonstationary**: the true values of the actions change over time. In this case there is a constant need for exploration.

## Exercises

### 2.2: Bandit example

Consider a k-armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using $\epsilon$-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the $\epsilon$ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

**Solution** The $\epsilon$ case definitely occured on $A_5$, because it was clear that prior to $A_5$, $a = 2$ was the greedy action choice. The $\epsilon$ case may have occured on steps $A_1, A_2$, because there were more than one greedy actions.

### 2.3: Cumulative reward in the long run

In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively.

**Solution** We first rule out the greedy method, as this method will only select the optimal choice about 1/3 of the time. Without exploration, it cannot improve in the long run.

The 0.01-greedy method would be better because in the long run, the 0.1-greedy method would only select the optimal action choice 90% of the time whereas the 0.01-greedy method would select the optimal choice 99% of the time.

Let us quantify this:

- on exploration moves, the methods pick the optimal action, $a^*$, with probability $\epsilon/n$,
- after convergence, the methods pick the optimal action with probability $1-\epsilon$.

Combined, this means that the methods will pick $a^*$ with probability $\epsilon/n + 1 - \epsilon$.

The expected reward for a given method can be expressed via the law of total expectation as follows:

$$\begin{aligned}\mathbb{E}(R_t) &= \sum_{a\in\mathcal{A}} \mathbb{E}(R_t|A_t=a)\mathbb{P}(A_t=a)\\ &= \mathbb{E}(R_t|A_t=a^*)\mathbb{P}(A_t=a^*) + \sum_{a\neq a^*} \mathbb{E}(R_t|A_t=a)\mathbb{P}(A_t=a)\\ &= \mathbb{E}(R_t|A_t=a^*)(\frac{\epsilon}{n}+1-\epsilon) + \sum_{a\neq a^*} \mathbb{E}(R_t|A_t=a)\frac{n-1}{n}\end{aligned}$$

We can take the limit as $n\to\infty$:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}(R_t) &= \mathbb{E}(R_t|A_t=a^*)(1-\epsilon) + \sum_{a\neq a^*} \mathbb{E}(R_t|A_t=a)\\ &= \mathbb{E}(R_t|A_t=a^*)(1-\epsilon) + 0 && \text{(since average reward of each run is 0)}\\ &= \begin{cases} 0.9\,\mathbb{E}(R_t|A_t=a^*), & \epsilon=0.1\\ 0.99\,\mathbb{E}(R_t|A_t=a^*), & \epsilon=0.01\end{cases}\end{aligned}$$

As shown, $\epsilon = 0.01$ yields better rewards in the long run.