Chapter 2: Multi-armed Bandits

2.1 A k-armed Bandit Problem

game rule

- k options, each has a random numerical reward
- **objective**: maximize expected total reward over time

2.2 Action-value Methods

Sample-average method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot 1_{A_i=a}}{\sum_{i=1}^{t-1} 1_{A_i=a}}$$

$Q_t(a)$ converges to $q_*(a)$ when the denominator goes to infinity.

- *greedy* action method

$$A_t \doteq \arg\max_x Q_t(a)$$

No exporation, only exploitation

- *ϵ-greedy* action method

  With probability ϵ, choose an action randomly equiprobably.

2.3 The 10-armed Testbed

ϵ-greedy betters off greedy when:

- reward variance is large
- reward is nonstationary

2.4 Incremental Implementation

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

The general form of it is:

$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize}[\text{Target} - \text{OldEstimate}]$$

The expression $[\text{Target} - \text{OldEstimate}]$ is an *error* in the estimate.

StepSize is denoted as α or $\alpha_t(a)$.

2.5 Tracking a Nonstationary Problem

If we take constant StepSize $\alpha \in (0, 1]$, we have:

$$Q_{n+1} \doteq Q_n + \alpha[R_n - Q_n]$$

And thus

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1-\alpha)Q_n \\ &= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\ &\dots \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \end{aligned}$$

In this case, Q_{n+1} is called an *exponential recency-weighted average* of past rewards and Q_1.

$\{\alpha_n(a)\}$ convergence condition

A well-known result in stochastic approximation theory gives us the conditions required to assure convergence with probability 1:

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \ \text{ and } \ \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

- The first condition is required to guarantee that the steps are large enough to eventually overcome any initial conditions or random fluctuations.
- The second condition guarantees that eventually the steps become small enough to assure convergence.

1. In sample average method, $\alpha_n(a) = \frac{1}{n}$ is bound to converge.
2. Constant $\alpha_n(a) = \alpha$ may not converge, but can respond to changes in nonstationary setup well. *Nonstationary problems are common in RL.*