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

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

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 non-stationary setup well. *Non-stationary problems are common in RL.*

2.6 Optimistic Initial Values

Initial Bias

Initial bias: The dependence of $Q_1(a)$

Choosing optimistic(high) initial values encourages exploration. All actions will be tried several times before converge.

This trick is effective on stationary problems but far from being a generally useful approach to encouraging exploration.

- In Non-stationary problems, its drive for exploration is temporary.

2.7 Upper-Confidence-Bound Action Selection

$$A_t(a) \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

$Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}$ is the upper-bound of $q_*(a)$ with confidence level c. This method takes uncertainty into consideration.

The requirement of storing $N_t(a)$ makes it impractical in large action space problems.

2.8 Gradient Bandit Algorithms

Another approach instead of $Q_t(a)$: learning a numerical *preference* function $H_t(a)$ overtime to determine the probability of choosing the action, according to *soft-max distribution*.

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

$H_t(a)$ can be updated using stochastic gradient ascent:

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)), && \text{and} \\ H_{t+1}(a) &\doteq H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a), && \text{for all } a \neq A_t \end{aligned}$$

$\bar{R}_t(a)$ is the average of all the rewards up through and including time t. It serves as the baseline of rewards. If current reward exceeds the baseline, the preference increases.

How to compute the gradient

We want to have

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

Where

$$\mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x)$$

Of course, it is not possible to implement gradient ascent exactly in our case because by assumption we do not know the $q_*(x)$, but in fact the updates of our algorithm are equal to the equation above in **expected value**, making the algorithm an instance of *stochastic gradient ascent*.

$$
\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)}\left(\sum_x \pi_t(x) q_*(x)\right) \\
&= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \sum_x \left(q_*(x) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)} && \Leftarrow \sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0
\end{aligned}
$$

B_t is the *baseline* mentioned above, in this case it is $\bar{R}_t$. We can now proceed and turn this equation into an expectation.

$$
\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \sum_x \left(q_*(x) - B_t\right) \pi_t(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x) \\
&= \mathbb{E}[(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t)] && \Leftarrow A_t \leftarrow x \\
&= \mathbb{E}[(R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t)] && \Leftarrow \mathbb{E}[R_t | A_t] = q_*(A_t)
\end{aligned}
$$

Now we take a look at the $\frac{\partial \pi_t(A_t)}{\partial H_t(a)}$.

$$
\begin{aligned}
\frac{\partial \pi_t(A_t)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)}\left(\frac{e^{H_t(A_t)}}{\sum_b e^{H_t(b)}}\right) \\
&= \begin{cases} -\pi_t(A_t)\pi_t(a), & A_t \neq a \\ \pi_t(A_t) - \pi_t(A_t)\pi_t(a), & A_t = a \end{cases} \\
&= \pi_t(A_t)(1_{A_t=a} - \pi_t(a))
\end{aligned}
$$

Finally

$$
\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \mathbb{E}[(R_t - \bar{R}_t)\pi_t(A_t)(1_{A_t=a} - \pi_t(a))/\pi_t(A_t)] \\
&= \mathbb{E}[(R_t - \bar{R}_t)(1_{A_t=a} - \pi_t(a))]
\end{aligned}
$$