Notes: Sutton Introduction to RL

Xingdong Zuo

IDSIA, Switzerland

March 15, 2018

1 (Ch. 2) Multi-armed Bandits

1.1 A K-armed Bandit Problem

Stationary K-armed bandit: Given $K \in \mathbb{N}^+$ possible actions associated with a set of stationary reward distributions $\{R_1, \ldots, R_K\}$. At each time step t, an action A_t is selected and a reward R_t is observed.

- Objective: maximize the expected total reward over T time steps
- Expected reward for action a: $Q^*(a) = \mathbb{E}[R_t | A_t = a]$
- Estimated action value: $Q_t(a) \approx Q^*(a)$

1.2 Action-value Methods

Theorem 1.1 (Law of large numbers)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu, \forall i = 1, 2, \ldots,$ then the sample average $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ converges to μ as $N \to \infty$.*

- Estimate action value by sample average:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}(A_i = a)}{\sum_{i=1}^{t-1} \mathbb{1}(A_i = a)} \tag{1}$$

- By law of large numbers, $Q_t(a)$ converges to $Q^*(a)$ as a being selected infinitely many times.
- Greedy action (exploitation): $A_t = \operatorname{argmax}_a Q_t(a)$
- ϵ-greedy (exploration): Sample $z \sim \mathcal{U}(0, 1)$

$$A_t = \begin{cases} \operatorname{argmax}_a Q_t(a) & \text{if } z < 1 - \epsilon \\ \mathcal{U}\{1, \ldots, K\} & \text{otherwise} \end{cases} \tag{2}$$

1.3 The 10-armed Testbed

- High uncertainty (large variance): more exploration
- No uncertainty (zero variance): greedy strategy is optimal

1.4 Incremental Implementation

We can estimate $Q_t(a)$ iteratively. Let R_i be the observed reward for i-th selection of action a, we have

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

General form of such update rule:

$$\text{NewEstimate} = \text{OldEstimate} + \text{StepSize}(\text{Target} - \text{OldEstimate}).$$

1.5 Tracking a Nonstationary Problem

Nonstationary problem: give more weight to recent rewards, e.g. constant step size $\alpha \in (0, 1]$, we have

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha\left(R_n - Q_n\right) \\
&= \alpha R_n + (1-\alpha)Q_n \\
&= \alpha R_n + (1-\alpha)\left(\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right) \\
&= \alpha R_n + \alpha(1-\alpha)R_{n-1} + \cdots + \alpha(1-\alpha)^{n-1}R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i
\end{aligned}$$

By applying geometric series, one can show

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1. \tag{3}$$

Thus, the update rule is a weighted average.

Adaptive step size $\alpha_n(a)$: the convergence condition is Monro-Robbins sequence, i.e.

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty \tag{4}$$

1.6 Optimistic Initial Values

Setting initial estimate $Q_1(a) > 0, \forall a$ encourages exploration, i.e. initially any selected action reduces its estimate, resulting in other actions to be considered. The exploration decreases over time.

1.7 Upper Confidence Bound Action Selection

- Problem of ϵ-greedy: treats non-greedy actions equally despite of estimation uncertainty.
- UCB action selection:

$$A_t = \underset{a}{\operatorname{argmax}} \left(Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right) \tag{5}$$

  where $N_t(a) = \sum_{i=1}^{t-1} \mathbb{1}(A_i = a)$ and the number $c > 0$ controls the degree of exploration.

  - Square root indicates uncertainty measure (variance)
  - Each time for selected action: uncertainty decreases
  - Each time for unselected action: uncertainty increases
  - Use of logarithm: increasing slower over time, but unbounded

1.8 Gradient Bandit Algorithms

- Action preference: $H_t(a)$
- Softmax policy:

$$\pi_t(a) = \mathrm{P}(A_t = a) = \frac{e^{H_t(a)}}{\sum_{b=1}^{K} e^{H_t(b)}} \tag{6}$$

- Objective:

$$\text{maximize } \mathbb{E}\left[R_t\right] = \sum_x \pi_t(x) Q^*(x) \tag{7}$$

- Update rule:

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)} \tag{8}$$

  for some $\alpha > 0$.

Lemma 1.2. $\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)\left(\mathbb{1}(a = x) - \pi_t(a)\right)$

Proof.

$$
\begin{aligned}
\frac{\partial \pi_t(x)}{\partial H_t(a)} &= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_{y=1}^K e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^K e^{H_t(y)}}{\partial H_t(a)}}{\left(\sum_{y=1}^K e^{H_t(y)}\right)^2} \\
&= \frac{\mathbb{1}(a=x) e^{H_t(x)}}{\sum_{y=1}^K e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{\left(\sum_{y=1}^K e^{H_t(y)}\right)^2} \\
&= \pi_t(x) \left(\mathbb{1}(a=x) - \pi_t(a)\right).
\end{aligned}
$$

□

By applying Lemma 1.2, we can obtain

$$
\begin{aligned}
\frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)} &= \sum_x Q^*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \sum_x Q^*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} - B_t \frac{\partial}{\partial H_t(a)} \sum_x \pi_t(x) \\
&= \sum_x \pi_t(x) \frac{1}{\pi_t(x)} \left(Q^*(x) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \mathbb{E}_{A_t \sim \pi_t(\cdot)} \left[\left(\mathbb{E}\left[R_t | A_t\right] - B_t\right) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \frac{1}{\pi_t(A_t)}\right] \\
&= \mathbb{E}_{A_t \sim \pi_t(\cdot)} \left[\left(R_t - B_t\right) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \frac{1}{\pi_t(A_t)}\right] \\
&= \mathbb{E}_{A_t \sim \pi_t(\cdot)} \left[\left(R_t - B_t\right) \left(\mathbb{1}(A_t = a) - \pi_t(a)\right)\right].
\end{aligned}
$$

We choose the baseline as averaged rewards prior to time t, i.e. $B_t = \bar{R}_t$, then we obtain the Monte-Carlo update rule

$$
H_{t+1}(a) = H_t(a) + \alpha \left(R_t - \bar{R}_t\right) \left(\mathbb{1}(A_t = a) - \pi_t(a)\right). \tag{9}
$$

for all $a \in \{1, \ldots, K\}$.