Multi-armed bandit and Markov decision process

by Jack Yansong Li

Liii Network

Email: yansong@liii.pro

Abstract

This short notes is a proof of the equivalence between Multi-armed bandit and finite state-action discounted Markov decision process

We first introduce the definition of multi-armed bandit:

Definition 1. *(Multi-armed bandit)* A Multi-Armed Bandit (MAB) is a tuple (A, r), where A is the set of actions/arms and

$$r: A \to \Delta(\mathbb{R}_+)$$

is a stochastic reward function that maps an action to a distribution over $\mathbb{R}_+ \triangleq [0, \infty)$.

By extending Definition 1 to multi-states, we have a Markov decision process.

Definition 2. *(Markov Decision Process)* A Markov Decision Process (MDP) is a tuple $(S, A, \mathbb{P}, r, \gamma)$, where $S \times A$ is the set of state-action pairs and

$$r: S \times A \to \Delta([0, 1])$$

is a stochastic reward function that maps a state-action pair to a distribution over $[0, 1]$. Similarly, the transition kernel $\mathbb{P}$ is a stochastic function that maps a state-action pair to a distribution over next state, formally defined as

$$\mathbb{P}: S \times A \to \Delta(S).$$

We now prove that a finite state and finite action MDP can be reduced to a MAB.

Theorem 1. *Consider a MDP defined in Definition 2 that satisfies $|S| < \infty$, $|A| < \infty$, and $\gamma < 1$. Then, the MDP $(S, A, r, \mathbb{P}, \gamma)$ can be reduced to a MAB.*

Proof. For any state-action pair (s, a), we can define expected cumulative reward r_{cum} as

$$r_{\text{cum}}(s, a) \triangleq \mathbb{E}_{s_{t+1} \sim \mathbb{P}(s_t, a_t)} \left(\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| s_1 = s, a_1 = a \right).$$

The limit for $\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)$ exists since $\gamma < 1$ and $r(s, a) \leq 1$. Therefor, the MDP $(S, A, r, \mathbb{P}, \gamma)$ can be reduced to a MAB with set of actions/arms be $(S \times A)$ and reward function be r_{cum}. $\square$