Notes on Reinforcement Learning

Gene Li

November 1, 2019

Abstract

These notes more or less follow the standard presentation found in Sutton and Barto. At times I have rearranged material to make it clearer for myself to understand.

1 Mult-Armed Bandits

The multi-armed bandit (MAB) problem represents a simple first stab at understanding reinforcement learning. The structure of MAB is simple in the sense that we do not need to worry about state transitions; each action that the player takes in a round proceeds independently in some sense.

First, we define the multi-armed bandit setup.

Definition 1.1 (Mult-Armed Bandit)**.** *Consider the following learning problem. At each round $t = 1, 2, ..., T$, the player selects one action among k choices, denoted $A_t \in [k]$ (typically called **arms**). They receive a reward $r_t \sim R_t(A_t)$, where R_t is some distribution which depends on the action a_t.*

*We define the **expected reward** of action a as:*

$$q_\star(a) := \mathbb{E}[R_t(A_t)|A_t = a]. \tag{1.1}$$

Roughly speaking, the goal of the player is to select the action(s) A_t which give the greatest reward R_t. However, a priori, the player does not know the distributions $R_t(a)$, so they must trade-off between learning these distributions (called *exploration*) and selecting the action which they believe gives the greatest reward (called *exploitation*).

1.1 Exploitation: Greedy and ϵ-Greedy Actions

1.2 Exploration: Estimating the Expected Reward