

## Reinforcement Learning – Sutton 2<sup>nd</sup> Edition

### Multiarmed Bandits

The most important feature distinguishing RL from other types of learning is that it uses training information that *evaluates* the actions taken rather than *instructs* by giving correct actions. This is what creates the need for active exploration, for an explicit search for good behavior. Purely evaluative feedback indicates how good the action taken was, but not whether it was the best or the worst action possible. Purely instructive feedback, on the other hand, indicates the correct action to take, independently of the action actually taken. This kind of feedback is the basis of supervised learning, which includes large parts of pattern classification, artificial neural networks, and system identification. In their pure forms, these two kinds of feedback are quite distinct: evaluative feedback depends entirely on the action taken, whereas instructive feedback is independent of the action taken.

In this Section we study the evaluative aspect of RL in a simplified setting, one that does not involve learning to act in more than one situation. This nonassociative setting is the one in which most prior work involving evaluative feedback has been done and it avoids much of the complexity of the full reinforcement learning problem. Studying this case allows us to see how evaluative feedback differs from, and yet can be combined with, instructive feedback.

The particular nonassociative, evaluative feedback problem that we explore is a simple version of the  $k$ -armed bandit problem.

#### A $k$ -armed Bandit Problem

Consider the following learning problem. You are faced repeatedly with a choice among  $k$  different options or actions. After each choice you receive a numerical reward chosen from a stationary probability distribution that depends on the action you selected. Your objective is to maximize the expected total reward over some time period, for example over 1,000 action selections or *time steps*.

This is the original form of the  $k$ -armed bandit problem, so named by analogy to a slot machine or “one-armed bandit”, except