Part I

Stochastic Optimization

Chapter 1

Finite Markov Decision Process

1.1 Introduction

Definition 1.1.1 (Markov Decision Process)**.** A (finite) Markov Decision Process is a 5-tuple $< \mathbb{S}, \mathbb{A}, \mathbb{T}, r, \gamma >$. In which

- $\mathbb{S}$ is a finite set of states
- $\mathbb{A}$ is a finite set of actions
- $\mathbb{T}$ is a state transition probability function

$$T(s'|s,a) = \mathbb{P}(S_{t+1} = s'|S_t = s, A_t = a) \tag{1.1}$$

- r is a reward function

$$r(s,a) = \mathbb{E}(R_{t+1}|S_t = s, A_t = a) \tag{1.2}$$

- γ is a discount factor $\gamma \in [0,1]$

Definition 1.1.2 (Return)**.** The **return** G_t is the total discounted reward from time-step t

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1.3}$$

Notice: The objective in Reinforce Learning is to maximize G_∞, that is, to choose A_t to maximize $R_{t+1}, R_{t+2}, \cdots$

Definition 1.1.3 (Policy)**.** A **policy** π is a distribution over actions given states.

$$\pi(a|s) = \mathbb{P}(A_t = a|S_t = s) \tag{1.4}$$

A policy fully defines the behavior of an agent.