Table of Contents:

- Preface
  - Introduction
    - Policy
    - Rewards
    - Objective
  - Markov Decision Process
    - Value Function
  - Dynamic Programming
    - Policy Iteration
- Introduction
- References:

Preface

Introduction

The essence of RL is learning through *interaction*. An RL agent interacts with its environment and, upon observing the consequences of its actions, can learn to alter its own behaviour in response to rewards received.[TODO: Add image for RL]

At minimal, the reinforcement learning problem has

- observation from environment, *state* s_t
- interaction between agent & environment, *action* a_t
- outcome of interation, *new state* s_{t+1}
- feedback of taking action in that state, *reward* r_{t+1}

Beginning from a state s_0, we take sequence of actions {$a_0, a_1, ...$} transitioning to subsequent states {$s_1, s_2, ...$} and collecting rewards on each transition {$r_1, r_2, ...$}.

Policy

The collection of actions can be abstracted as a *policy* π_t that maps from states to probablities $\pi_t(a|s)$ of selecting each possible action in that state.

Rewards

For each policy, the agents collects rewards and this cumulative reward measures the quality of the policy. Simply, it can be total sum of rewards or a more popular *discounted reward* that assigns weighs to future rewards.

So, if γ is the discount rate denoting the importance of immediate rewards versus the future rewards, the accumulated return can be formalized as,

$$G_t(s) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

where \(0 \leq \gamma \leq 1 \).

- \(\gamma=0\), agent is myopic & maximizes only the immediate reward.
- \(\gamma=1\), is simply sum of all rewards.
- \(0 < \gamma < 1 \), farsighted & values future rewards but not as much as immediate rewards.

For each policy, the total rewards expected by the agent to collect is known as *Expected Reward/Return*.

Objective

The goal of the agent is to find an optimal policy \(\pi^* \) that maximizes the *expected return* in the environment.

Markov Decision Process

What is Markov Property ? How is MDP formulated ? What is a value function ?

Value Function

Dynamic Programming

Policy Iteration

Introduction

- What is Maonte Carlo Tree Search Method ?
- Why MCTS ?
-

References:

1. A brief Survey of Deep Reinforcement Learning
2. Reinforcement Learning: An Introduction, Sutton & Barto