COMP90054 – Week 8 tutorial

Last updated: 24 April 2023

Markov Decision Process (MDP)

Definition

- A MDP is a fully observable, probabilistic state model to specify a problem
  - c.f. State-space model in classical planning
  - Model → Solver → Solution
- In this subject, we focus on discount-reward MDPs

$$P = \langle S, s_0, A, P, r, \gamma \rangle$$

S	State space S, finite and discrete
s_0	A **known** Initial state $s_0 \in S$
A	A set of actions, with $A(s) \subseteq A$ for each $s \in S$
$P_a(s'\|s)$	**Transition probabilities** for $s \in S$ and $a \in A(s)$
$r(s, a, s')$	**Reward** for transitioning from state s to state s' with action a
γ	**Discount factor** $0 \leq \gamma < 1$

- For discounted-reward MDPs, optimal solutions maximise the expected discounted accumulated reward from the initial state

Policy

- Solution to a MDP is called policy π, a function that tells an agent which action is the best one to choose in each state
- A policy can be
  - deterministic $\pi: S \to A$, which maps each state to one best action
  - stochastic $\pi: S \times A \to \mathbb{R}$, which specifies the probability distribution from which the agent should select an action

Solver for MDP

- In this subject we focus on two techniques for solving MDPs:
  - Value-based methods: learn the value of states and actions, then extract a policy
  - Policy-based methods: learn the policy directly

Value Iteration

- Value iteration is a value-based algorithm for finding the optimal value function V^* by solving Bellman equations iteratively
  - Value function $V: S \rightarrow \mathbb{R}$ assigns each state with a value

Idea

- For each iteration over states and actions, update the Q table until it converges

$$Q(s,a) = \sum_{s' \in S} P_a(s'|s) \times [r(s,a,s') + \gamma \times V(s')]$$

	Action1	Action2
State1	a	b
State2	c	d

- Find the value of a state using $V(s) = \max_{a \in A(s)} Q(s,a)$
- Extract the policy by $\pi(s) = \underset{a \in A(s)}{\operatorname{argmax}} Q(s,a)$
  - In a state s, given V, choose the action with the highest expected reward