MLSEC
Reinforcement Learning
Draft v.2

Characteristics of RL

- There is no supervisor, only a reward
- Feedback is delayed
- Time really matters (sequential data)
- Agent's actions affect the subsequent data it receives

Rewards

- A reward R_t is a scalar feedback signal.
- It indicates how well agent is doing at step t.
- The agent's job is to maximise a cumulative reward.

Reward Hypothesis

All goals can be described by the maximisation of expected cumulative reward.

Sequential Decision Making

- Goal: select actions to maximise total future reward
- Actions may have long term consequences
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward

Examples

- A financial investment (may take months to mature)
- Refuelling a helicopter (might prevent a crash in several hours)
- Blocking opponent moves (might help winning chances many moves from now)

Markov Decision Process

Definition

A Markov decision process is a 4-tuple (S, A, P, R), where:

- S is a finite set of states called the state space;
- A is a finte set of actions called the action space; for $s \in S$ let A_s be the set of all actions available from s; $A = \bigcup_{s \in S} A_s$; $A(S) = \{(s, a) \mid s \in S, a \in A_s\}$;
- $P : A(S) \times S \to [0, 1], (s, a, s') \mapsto P_a(s, s')$ is a function, where $P_a(s, s') = \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action a will transform state s at time t into state s' at time $t + 1$;
- $R : A(S) \to \mathbb{R}, (s, a) \mapsto R_a(s)$ is a function, where $R_a(s)$ is the immediate reward received after transitioning from state s due to action a.

MDP Example: Student's environment (states and actions)

$S = \{\text{Class 1, Class 2, Class 3, Pass, Sleep, Pub, Facebook}\}$
$A = \{\text{Study, Go to FB, Stay, Quit, Fall asleep, Go to pub,}$
$\text{Go back to 1, Go back to 2, Go back to 3}\}$
$A_{\text{Class 1}} = \{\text{Study, Go to FB}\}$
$A_{\text{Class 2}} = \{\text{Study, Fall asleep}\}$
$A_{\text{Class 3}} = \{\text{Study, Go to pub}\}$
$A_{\text{Pass}} = \{\text{Fall asleep}\}$
$A_{\text{Sleep}} = \emptyset$
$A_{\text{Pub}} = \{\text{Go back to 1, Go back to 2, Go back to 3}\}$
$A_{\text{Facebook}} = \{\text{Stay, Quit}\}$

MDP Example: Student's environment (Rewards)

Policies

Definition

A (deterministic) policy is a map $\pi : S \to A$ such that $\pi(s) \in A_s$ for all $s \in S$.

Definition

A (stochastic) policy is a map $\pi : S \to \mathcal{M}^1(A)$ such that $\pi(s) \in \mathcal{M}^1(A_s)$ for all $s \in S$.

Remarks

- $\mathcal{M}^1(A) = \{p \mid p : A \to [0,1], \sum_{a \in A} p(a) = 1 \ \}$
- We write $\pi(a \mid s)$ for $\pi(s)(a)$.

Rewards, goals and values I

Let $s_1, a_1, ..., a_{T-1}, s_T$ be a sequence of states and actions such that $a_t \in A(s_t)$ and

$$P_{a_t}(s_t, s_{t+1}) > 0.$$

for all t. Put

$$G = R_{a_1}(s_1) + ... + R_{a_{T-1}}(s_{T-1}).$$

Rewards, goals and values II

Let $s_1, a_1, ..., a_{T-1}, s_T$ be a sequence of states and actions as above and $\gamma \in [0, 1)$.Put

$$G = \sum_j \gamma^j R_{a_j}(s_j)$$

Rewards, goals and values III

Let π be a policy w.r.t. S and A.

Definition (State-value function)

For $s \in S$ define

$$v_\pi(s) = \mathbb{E}_\pi[G \mid s_1 = s]$$

The function $v_\pi : S \to \mathbb{R}$ is called the state-value function.

Theorem (Bellman equation)

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \left(R_a(s) + \gamma \sum_{s' \in S} P_a(s, s') v_\pi(s') \right)$$

Definition

$\pi \leq \pi'$ if $v_\pi(s) \leq v_{\pi'}(s)$ for all $s \in S$.

Rewards, goals and values IV

Definition (Action-value function)

For $s \in S, a \in A(s)$ define

$$q_\pi(s, a) = \mathbb{E}_\pi[G \mid s_1 = s, a_1 = a]$$

The function $q_\pi : A_S \to \mathbb{R}$ is called the action-value function.

Theorem

$$q_\pi(s, a) = R_a(s) + \gamma \sum_{s' \in S} P_a(s, s') v_\pi(s')$$

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) q_\pi(s, a)$$

Bellman eqution as a fixed point relation

Let $V = \{v \mid v : S \to \mathbb{R}\}$ and

$$T_\pi : V \to V$$

defined by

$$T_\pi v(s) = \sum_{a \in A(s)} \pi(a \mid s) \left(R_a(s) + \gamma \sum_{s' \in S} P_a(s, s') v(s') \right).$$

Then

$$v_\pi = T_\pi v_\pi.$$

Properties of T_π

T_π is a contraction:

$$T_\pi v(s) - T_\pi w(s) = \gamma \sum_{a \in A(s)} \pi(a \mid s) \sum_{s' \in S} P_a(s,s') \left(v(s') - w(s')\right),$$

thus

$$\max_s |T_\pi v(s) - T_\pi w(s)| \le \gamma \max_s |v(s) - w(s)|.$$

Banach's fixed point theorem shows that there is a unique fixed point of T_π and for each v the sequence $(T_\pi)^j v$ converges to this fixed point.