UMassAmherst
The Commonwealth's Flagship Campus

Markov decision processes- Terminology

Chaitra Gopalappa

Reference

- Chapter 3, Sutton and Barto,
- https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

Anatomy of general MDP formulation

- Let X_t be the system state at time t
- Let D_t be the decision at time t
- $\{X_t, D_t\}|_{t=0}^{\infty}$ is a Markov decision process defined by the n-tuple $\{S, A, P_a, R_a\}$
  - S is the state space (observation space in RL);
  - A is the action space
  - P_a is the transition probability matrix corresponding to an action 'a'
    - An element $p(i, a, j) =$ probability of transitioning to state j when system is in state i and action a is taken
  - R_a is the immediate reward matrix corresponding to an action 'a'
    - An element $r(i, a, j) =$ immediate reward of taking action a when system is in state i and transitioning to state j
- A policy (π) is a vector of size |S|, referring to the action to be taken in corresponding state

Objective function of MDP- Two broad types

- Two broad types of objective function
  - State−value function
    - $v_*(s) \sim v_{\pi^*}(s) = \max_{\pi} v_\pi(s), \forall s \in S$
    - $\pi^* = arg \max_{\pi} v_\pi(s), \forall s \in S$
    - $v_\pi(s)$: value function of state s under some policy π
    - It is the **expected** reward when **starting** in state s and following some policy π
    - $v_*(s)$: value function of state s under optimal policy π^*
  - Action−value function
    - Q-value: We will discuss this in RL

Rewards (immediate reward) v. Returns

- Reward: $r(i, a, j)$ = immediate reward of taking action a when system is in state i and transitions to j
- *Returns at time t*
  - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
  - γ: discount factor
  - We can set $\gamma = 1$ (when can we do this?)

Value function is the *expected* returns

- $v_\pi(s) =$ The value function of a state s under policy π
- Three types of value functions
  - (expected average returns) $v_\pi(s) = \mathbb{E}_\pi \frac{[\sum_{k=0}^{T} R_{t+k+1} | S_t = s]}{T}; \forall s \in S$
  - (expected discounted total returns) $v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\ [\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s]]; \forall s \in S$
  - (expected undiscounted total returns) $v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\ \left[\sum_{k=0}^{T} R_{t+k+1} | S_t = s\right]; \forall s \in S$

- Note: $v_\pi(s)$ is the value of s implies that it is the total expected returns of policy π if we start at s at time t (i.e., $S_t = s$)and follow policy π from there on

When to use value function=Expected "average" returns

- (expected average returns) $v_\pi(s) = \mathbb{E}_\pi \frac{[\sum_{k=0}^{T} R_{t+k+1} | S_t = s]}{T}; \forall s \in S$
  - $v_\pi(s) = \sum_{s \in S} \rho_\pi(s)\, \bar{r}(s, \pi(s))$
  - ρ_π **is the steady state distribution under policy** $\boldsymbol{\pi}$
  - $\bar{r}(s, \pi(s)) \sim \bar{r}(s, \pi(s) = a)$
  - $\bar{r}(i, a) = \sum_{j \in S} r(i, a, j) p(i, a, j)$
  - Note: if exhaustive enumeration, we calculate this for all policies; and the optimal policy is one with highest value
    - Number of possible policies = $|A|^{|S|}$
- Expected "average" returns: Most suitable for systems that are continuous
  - Inventory control, production planning
  - Can be represented by regular Markov chains (has a steady state)

- Would this work for systems that have a terminating state?
  - Games (win/lose); epidemics (eliminate/wipeout); robotic tasks;

When to use value function= expected "total" returns (general expression)

- $v_\pi(s)$ = The value function of a state s under policy π
  - (expected discounted total returns)
  - $v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right]; \forall s \in S$$

General expression that works in all cases

- *This works for both **continuous (steady state) and episodic (terminating state)** systems*
- $\gamma = 1$ if we want undiscounted
  - If $\gamma < 1$, future rewards get lower value
- t can be zero (or any other number); what the expression is saying is, if we are in state s and follow policy π the value function of s can be expressed as above.

Notational variations

Legend:
~ different notations to indicate the same metric

- Transition probabilities

1. $P_a[i,j] \sim p(i,a,j) \sim p(j|i,a)$
   - Used when referring to transition probability specific to an action
2. $p(S_t = s, \pi(s), S_{t+1} = s') \sim p(s, \pi(s), s') \sim p(s'|s, \pi(s))$
   - $\pi(s)$: action corresponding to sate s when following policy π
   - $S_t = s$: state at time 't' is s.
   - Used when referring to transition probability at time t when following some policy π; state at time t is s, and system transitions to state s' upon taking action $\pi(s)$

- Immediate rewards

1. $R_a[i,j] \sim r(i,a,j) \sim r(j|i,a)$
   - Used when referring to immediate reward specific to an action
2. $R_t \sim r(S_t = s, \pi(s), S_{t+1} = s') \sim r(s, \pi(s), s') \sim r(s'|s, \pi(s))$
   - Used when referring to immediate reward at time t when following some policy π; state at time t is s, and system transitions to state s' upon taking action $\pi(s)$

Solution algorithms to solve MDPs?

- Exhaustive enumeration
- Dynamic Programming
- Reinforcement learning
- Deep reinforcement learning

UMassAmherst

The Commonwealth's Flagship Campus