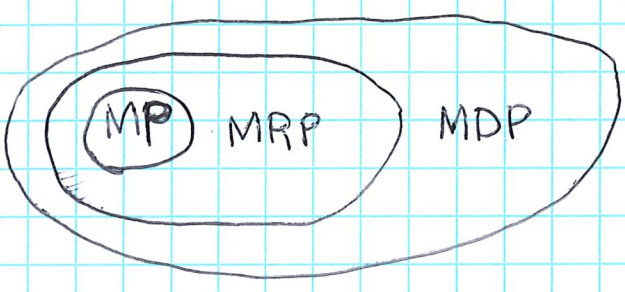
Markov Decision Processes



MP - Markov Process
MRP - Markov Reward Process
MDP - Markov Decision Process

MP

- S - states
- P - state transition matrix : $P_{s,s'} = \mathbb{P}(S_{t+1}=s' \mid S_t = s)$

MRP

- R - reward function $\mathcal{R}: S \to \mathbb{R}$ $\mathcal{R}_s = \mathbb{E}[R_{t+1} \mid S_t = s]$

(note difference btwn $\mathcal{R}$ & R_i)

- γ - discount factor $\gamma \in [0, 1)$

MDP

- A - set of actions
- $\mathcal{R}$ - reward fcn, depends on a: $\mathcal{R}_{s,a} = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- P - state trans. matrix, fcn. of a: $P^a_{s,s'} = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$

MRP

State-value function: $v(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 \cdot R_{t+3} + \dots \mid S_t = s]$

Expected discounted Rewards from s

$= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} \mid S_t = s\right]$

$$\underline{v} = \begin{bmatrix} v(s_1) \\ \vdots \\ v(s_m) \end{bmatrix}$$

$$\underline{v} = \mathcal{R} + \gamma \cdot P \underline{v}$$

$$\underline{v} = (I - \gamma P)^{-1} \cdot \mathcal{R}$$ (algebra)

MDP

Policy - $\pi(a|s) = \mathbb{P}(A_t = a \mid S_t = s)$ ← "Strategy"

For a fixed MDP & policy:

→ $\{S_i\}_{i=1}^{\infty}$ is a MP $\quad \{(S_i, R_i)\}_{i=1}^{\infty}$ is a MRP

Ex: Given a policy π & MDP, how to compute transition matrix $P_{s,s'}$?

$$P_{s,s'} = \sum_{a \in A} \pi(a|s) P^a_{s,s'}$$

State-Value Function: $v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R^{(\pi)}_{t+k+1} \mid S_t = s\right]$

$= \mathbb{E}_\pi\left[R^{(\pi)}_{t+1} + \gamma \cdot v_\pi(S_{t+1}) \mid S_t = s\right]$

$$v_\pi = (I - \gamma \cdot P^{(\pi)})^{-1} \cdot \mathcal{R}^{(\pi)}$$

Action-Value Function: $q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot R^{(\pi)}_{t+k+1} \mid S_t = s, A_t = a\right]$

Q: How to optimize Rewards?

Bellman Equations

$$v^*(s) = \max_{\pi} v_\pi(s)$$

$$q^*(s,a) = \max_{\pi} q_\pi(s,a)$$

1) $$v^*(s) = \max_a R_{s,a} + \gamma \cdot \sum_{s' \in S} P^a_{ss'} \cdot v^*(s')$$

$$q^*(s,a) = R_{s,a} + \gamma \sum_{s' \in S} P^a_{s,s'} \max_{a'} q^*(s',a')$$

$$\Downarrow$$

$$\pi^*(s) = \arg\max_a q^*(s,a)$$

Solve w/ iterative methods

- Value Iteration
- Q-learning