

# RL: Basics

## The Markov Decision Process

Marius Lindauer



Winter Term 2021

# Markov Decision Process (MDP)

- ▶ Markov Decision Process is Markov Reward Process + actions
- ▶ Definition of MDP
  - ▶  $S$  is a (finite) set of Markov states  $s \in S$
  - ▶  $A$  is a (finite) set of actions  $a \in A$
  - ▶  $P$  is dynamics/transition model for each action, that specifies  $P(s_{t+1} = s' \mid s_t = s, a_t = a)$
  - ▶  $R$  is a reward function  $R(s_t = s, a_t = a) = \mathbb{E}[r_t \mid s_t = s, a_t = a]$ 
    - ▶ Sometimes  $R$  is also defined based on  $(s)$  or on  $(s, a, s')$
  - ▶ Discount factor  $\gamma \in [0, 1]$
- ▶ MDP is tuple  $(S, A, P, R, \gamma)$

## MDP (cont'd)

- ▶ MDP is tuple  $(S, A, P, R, \gamma)$
- ▶ Optional components:
  - ▶  $\rho_0 : S \rightarrow \mathbb{R}^+$ : a distribution of start states
    - ▶ uniform distribution: the agent can start in any state – implicit assumption of MDP definition above
    - ▶ non-uniform distribution: the agent starts its episodes in only some of the states; e.g., it's unlikely that a game will start in a terminal state


## MDP (cont'd)

- ▶ MDP is tuple  $(S, A, P, R, \gamma)$
- ▶ Optional components:
  - ▶  $\rho_0 : S \rightarrow \mathbb{R}^+$ : a distribution of start states
    - ▶ uniform distribution: the agent can start in any state – implicit assumption of MDP definition above
    - ▶ non-uniform distribution: the agent starts its episodes in only some of the states; e.g., it's unlikely that a game will start in a terminal state
  - ▶  $T \subset S$ : set of terminal states
    - ▶ important for episodic MDPs
    - ▶ or if there is not fixed horizon, but the episodes should be finite

## MDP (cont'd)

- ▶ MDP is tuple  $(S, A, P, R, \gamma)$
- ▶ Optional components:
  - ▶  $\rho_0 : S \rightarrow \mathbb{R}^+$ : a distribution of start states
    - ▶ uniform distribution: the agent can start in any state – implicit assumption of MDP definition above
    - ▶ non-uniform distribution: the agent starts its episodes in only some of the states; e.g., it's unlikely that a game will start in a terminal state
  - ▶  $T \subset S$ : set of terminal states
    - ▶ important for episodic MDPs
    - ▶ or if there is not fixed horizon, but the episodes should be finite
  - ▶  $\gamma$ : discount factor
    - ▶ important to quantify the importance of future
    - ▶ some treat  $\gamma$  as a hyperparameter and not part of the definition
    - ~> different optimal policies can be found
    - ~> depends on how the optimal policy is defined

# Mars Rover as MDP

$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$
						

- ▶ 2 deterministic Actions: TryLeft and TryRight

# MDP Policies

- ▶ Policy specifies what action to take in each state
  - ▶ Can be deterministic or stochastic
- ▶ For generality, consider as a conditional distribution
  - ▶ Given a state, specifies a distribution over actions
- ▶ Policy:  $\pi(a \mid s) = P(a_t = a | s_t = s)$

# MDP + Policy

- ▶ MDP + Policy  $\pi(a \mid s)$  = Markov Reward Process
- ▶ Precisely, it is the MRP  $(S, R^\pi, P^\pi, \gamma)$  where

$$R^\pi(s) = \sum_{a \in A} \pi(a \mid s) R(s, a)$$

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

- ▶ Implies we can use same techniques to evaluate the value of a policy for an MDP as we could to compute the value of a MRP, by defining a MRP with  $R^\pi$  and  $P^\pi$



# MDP Policy Evaluation, Iterative Algorithm

- ▶ Goal: For a given  $\pi$ , determine  $V^\pi$
- ▶ iterative approach:
  - ▶ Initialize  $V_0(s) = 0$  for all  $s$
  - ▶ For  $k = 1$  until convergence
    - ▶ For all  $s$  in  $S$ :

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

- ▶ This is a Bellman backup for a particular policy

## Practice: MDP 1 Iteration of Policy Evaluation, Mars Rover Example

- ▶ Dynamics:  $p(s_6|s_6, a_1) = 0.5, p(s_7|s_6, a_1) = 0.5, \dots$
- ▶ Reward: for all actions, +1 in state  $s_1$ , +10 in state  $s_7$ , 0 otherwise
- ▶ Let  $\pi(s) = a_1. \forall s$ , assume  $V_k^\pi = [1, 0, 0, 0, 0, 0, 10]$  and  $k = 1, \gamma = 0.5$

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

## Practice: MDP 1 Iteration of Policy Evaluation, Mars Rover Example

- ▶ Dynamics:  $p(s_6|s_6, a_1) = 0.5, p(s_7|s_6, a_1) = 0.5, \dots$
- ▶ Reward: for all actions, +1 in state  $s_1$ , +10 in state  $s_7$ , 0 otherwise
- ▶ Let  $\pi(s) = a_1. \forall s$ , assume  $V_k^\pi = [1, 0, 0, 0, 0, 0, 10]$  and  $k = 1, \gamma = 0.5$

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

$$V_{k+1}^\pi(s_6) = 0 + \gamma[p(s_6 | s_6, a_1) \cdot V_k^\pi(s_6) + p(s_7 | s_6, a_1) \cdot V_k^\pi(s_7)]$$

## Practice: MDP 1 Iteration of Policy Evaluation, Mars Rover Example

- ▶ Dynamics:  $p(s_6 | s_6, a_1) = 0.5, p(s_7 | s_6, a_1) = 0.5, \dots$
- ▶ Reward: for all actions, +1 in state  $s_1$ , +10 in state  $s_7$ , 0 otherwise
- ▶ Let  $\pi(s) = a_1. \forall s$ , assume  $V_k^\pi = [1, 0, 0, 0, 0, 0, 10]$  and  $k = 1, \gamma = 0.5$

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

$$\begin{aligned} V_{k+1}^\pi(s_6) &= 0 + \gamma[p(s_6 | s_6, a_1) \cdot V_k^\pi(s_6) + p(s_7 | s_6, a_1) \cdot V_k^\pi(s_7)] \\ &= \gamma[0.5 \cdot 0 + 0.5 \cdot 10] \end{aligned}$$

## Practice: MDP 1 Iteration of Policy Evaluation, Mars Rover Example

- Dynamics:  $p(s_6|s_6, a_1) = 0.5, p(s_7|s_6, a_1) = 0.5, \dots$
- Reward: for all actions, +1 in state  $s_1$ , +10 in state  $s_7$ , 0 otherwise
- Let  $\pi(s) = a_1. \forall s$ , assume  $V_k^\pi = [1, 0, 0, 0, 0, 0, 10]$  and  $k = 1, \gamma = 0.5$

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

$$\begin{aligned} V_{k+1}^\pi(s_6) &= 0 + \gamma[p(s_6 | s_6, a_1) \cdot V_k^\pi(s_6) + p(s_7 | s_6, a_1) \cdot V_k^\pi(s_7)] \\ &= \gamma[0.5 \cdot 0 + 0.5 \cdot 10] \\ &= \gamma \cdot 5 \end{aligned}$$

## Practice: MDP 1 Iteration of Policy Evaluation, Mars Rover Example

- ▶ Dynamics:  $p(s_6|s_6, a_1) = 0.5, p(s_7|s_6, a_1) = 0.5, \dots$
- ▶ Reward: for all actions, +1 in state  $s_1$  , +10 in state  $s_7$  , 0 otherwise
- ▶ Let  $\pi(s) = a_1. \forall s$ , assume  $V_k^\pi = [1, 0, 0, 0, 0, 0, 10]$  and  $k = 1, \gamma = 0.5$

$$V_k^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' | s, \pi(s)) V_{k-1}^\pi(s')$$

$$\begin{aligned} V_{k+1}^\pi(s_6) &= 0 + \gamma[p(s_6 | s_6, a_1) \cdot V_k^\pi(s_6) + p(s_7 | s_6, a_1) \cdot V_k^\pi(s_7)] \\ &= \gamma[0.5 \cdot 0 + 0.5 \cdot 10] \\ &= \gamma \cdot 5 \\ &= 2.5 \end{aligned}$$

(1)

# MDP Control

- ▶ Compute the optimal policy

$$\pi^*(s) \in \arg \max_{\pi} V^{\pi}(s)$$

- ▶ There **exists a unique optimal value function**
- ▶ Optimal policy for an MDP in an infinite horizon problem is (i.e. agents acts forever is)
  - ▶ deterministic
  - ▶ stationary (does not depend on time step)
  - ▶ Unique?  $\rightsquigarrow$  Not necessarily, may have state-actions with identical optimal values