Reinforcement learning and robot navigation using MDPs

Charles Dufour

March 25, 2018

Introduction

The problem

- Framework : the Disopt robot which can follow lines
- The problem : the robot should adapt its speed with respect to traffic lights
- How : using Markov Decision Process (MDP)

MDPs

Definition

A Markov Decision Process (MDP) is a discrete time stochastic control process, used in situations where outcomes are and random and partly under the control of a decision maker.

MDPs

Definition (suite)

- A set of states $\mathcal{S} = \{s_0, s_1, s_2, \ldots\}$
- A set of actions $\mathcal{A} = \{a_1, a_2, a_3, \ldots\}$
- A transition function $T(a, s, s', r) = \mathbb{P}[s', r \mid a, s]$
- A reward function $R : \mathcal{S} \mapsto \mathbb{R}$
- A discount factor γ

Markov Property

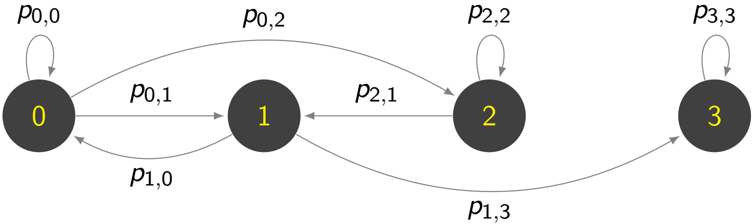
The transitions only depends on the current state and the current action.

to be told

~~Particularly, at each time step, our process is in some state s.~~
~~Then our learning agent decides which action to execute from the set $\mathcal{A}$ which is doable from state s.~~
~~Then the process moves randomly to a new state s' following T and gives the agent a reward $R(s')$.~~
~~The purpose of our agent is to maximise the cumulative reward it gets in the long run.~~

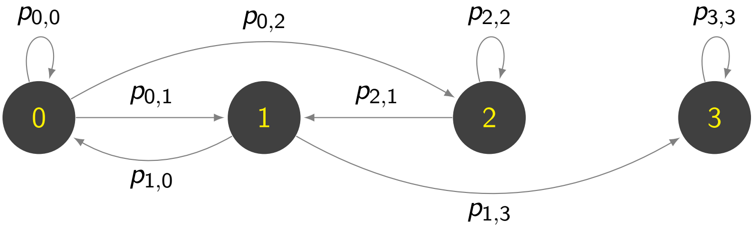
MDP example

MDP's can be easily represented by graphs :



MDP example

MDP's can be easily represented by graphs :



The constraints are $\sum_j p_{i,j} = 1 \quad \forall i \in \mathcal{S}$

how to pick actions

Definition

A policy π is a probabilistic mapping from the set of states to the set of actions :

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

Issue

How to ?

How to asses the goodness of policies so we can find the best one ?
What is the best policy ?

how to asses the goodness of policies

Discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k * R_{t+k+1}$$

how to asses the goodness of policies

Discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k * R_{t+k+1}$$

action value while in a state s under π

$$q_\pi(s, a) = \mathbb{E}[G_t \mid S_t = s, A_t = a] \tag{1}$$

how to asses the goodness of policies

Discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k * R_{t+k+1}$$

action value while in a state s under π

$$q_\pi(s, a) = \mathbb{E}[G_t \mid S_t = s, A_t = a] \tag{1}$$

state value under policy π

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t \mid S_t = s] \\ &= \sum_a \pi(a \mid s) \sum_{r,s'} p(s', r \mid s, a)[r + \gamma v_\pi(s')] \end{aligned} \tag{2}$$

how to asses the goodness of policies

how to compare two policies

$$\pi \leq \pi' \iff \pi(s) \leq \pi'(s) \quad \forall s \in \mathcal{S}$$

how to asses the goodness of policies

how to compare two policies

$$\pi \leq \pi' \iff \pi(s) \leq \pi'(s) \quad \forall s \in \mathcal{S}$$

Optimal policy

$$\pi_* \quad s.t. \quad \forall \pi : \pi_* \geq \pi$$

Bellman optimality equations

The optimal policy π_* has value functions : v_* and q_*

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')] \tag{3}$$

$$q_*(s, a) = \sum_{s',r} p(s', r \mid s, a)[r + \gamma \max_{a'} q_*(s', a')] \tag{4}$$

~~Intuitively these equations say that the value of a state under the optimal policy must equal the expected return for the best action from that state. For finite MDPs these equations have a unique solution.~~

Another issue

computational issue

If we wanted to solve these equations directly, it would cost a lot of computational power to know exactly the value functions first and then to solve. So how do we do it ?

Another issue

computational issue

If we wanted to solve these equations directly, it would cost a lot of computational power to know exactly the value functions first and then to solve. So how do we do it ?

Approximation of value function

solving MDPs using dynamic programming

policy iteration

update rule :

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)(r + \gamma v_k(s'))$$

solving MDPs using dynamic programming

policy iteration

update rule :

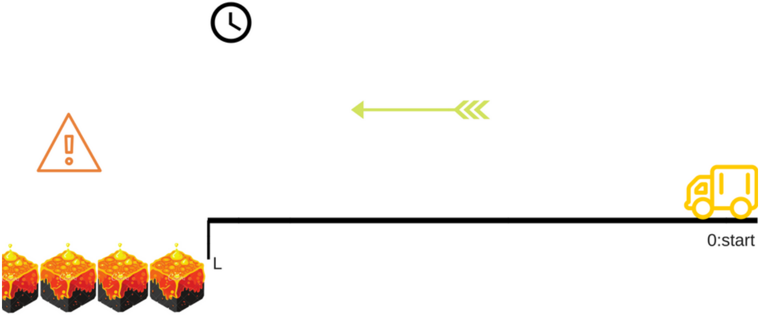
$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)(r + \gamma v_k(s'))$$

Policy Improvement

π / π' : old/new policy.

$$\pi'(s) = argmax_{a \in \mathcal{A}} q_{\pi}(s, a)$$

what have we done so far



modelization

States

modelization

States

- position $\{0,1,2,\ldots,L,\ \text{Lava}\ \}$

modelization

States

- position $\{$0,1,2,...,L, Lava $\}$
- speed $\{$low, medium, high $\}$

modelization

States

- position $\{$0,1,2,. . . ,L, Lava $\}$
- speed $\{$low, medium, high $\}$

Actions

modelization

States

- position $\{0,1,2,\ldots,L, \text{Lava}\}$
- speed $\{\text{low, medium, high}\}$

Actions

- decelerating

modelization

States

- position {0,1,2,. . . ,L, Lava }
- speed {low, medium, high }

Actions

- decelerating
- maintaining speed

modelization

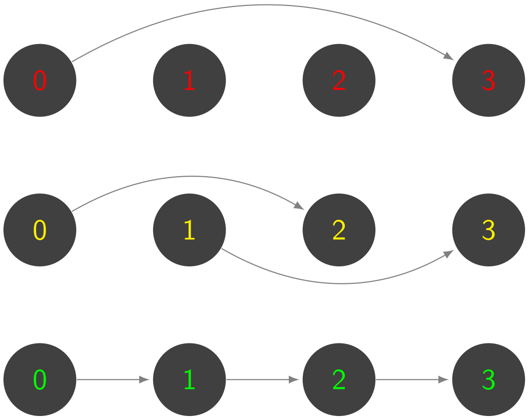
States

- position $\{0,1,2,\ldots,L, \text{Lava}\}$
- speed $\{\text{low, medium, high}\}$

Actions

- decelerating
- maintaining speed
- accelerating

accelerating graph

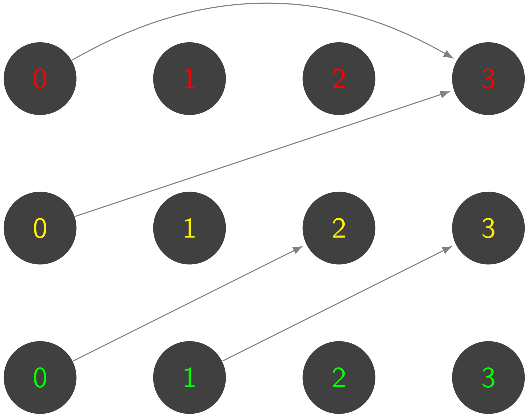


red : high speed

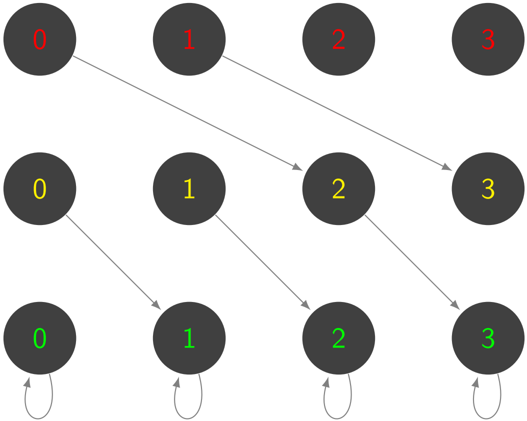
yellow : medium speed

green : low speed

keeping the same speed graph



decelerating



Results

What's next ?

already working on

Other ideas