



AIL 7022: Reinforcement Learning

Lecture 5: Formalising MDPs

Instructor: Raunak Bhattacharyya



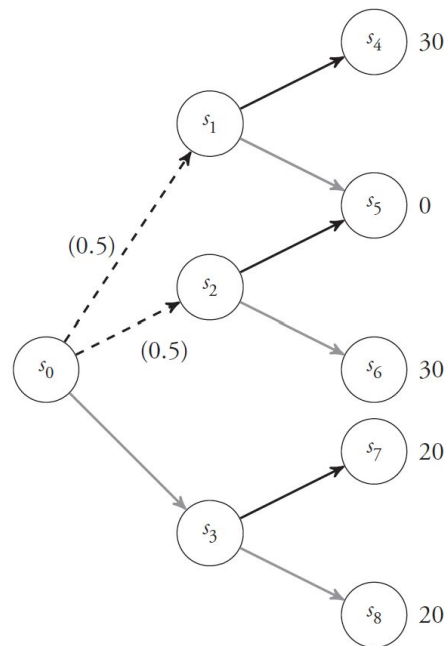
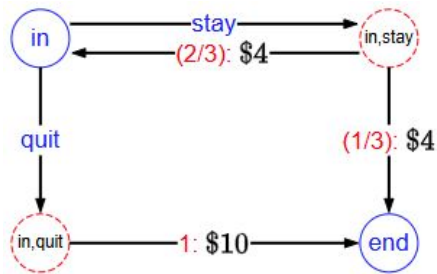
ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap



Source: [Dreamstime](https://www.dreamstime.com/stock-illustration-hot-dog-stand-vector-image-imagebank-gettyimages-image219996445.html)



$$J(\theta) = \mathbb{E}_{(s_1, a_1, s_2, a_2, \dots, s_T, a_T) \sim p_\theta(s_1, a_1, \dots, s_T, a_T)} \left[\sum_{t=1}^T r(s_t, a_t) \right]$$

Plan

- Formalising MDPs
- Value functions
- Policy Evaluation

Formalising MDPs

MDPs: Definition

Definition 2.1 **Finite Discrete-Time Fully Observable Markov Decision Process (MDP).** A finite discrete-time fully observable MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{D}, \mathcal{T}, \mathcal{R} \rangle$, where:

- $\mathcal{S}$ is the finite set of all possible states of the system, also called the *state space*;
- $\mathcal{A}$ is the finite set of all actions an agent can take;
- $\mathcal{D}$ is a finite or infinite sequence of the natural numbers of the form $(1, 2, 3, \dots, T_{max})$ or $(1, 2, 3, \dots)$ respectively, denoting the decision epochs, also called *time steps*, at which actions need to be taken;
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{D} \rightarrow [0, 1]$ is a transition function, a mapping specifying the probability $\mathcal{T}(s_1, a, s_2, t)$ of going to state s_2 if action a is executed when the agent is in state s_1 at time step t ;
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{D} \rightarrow \mathbb{R}$ is a reward function that gives a finite numeric reward value $\mathcal{R}(s_1, a, s_2, t)$ obtained when the system goes from state s_1 to state s_2 as a result of executing action a at time step t .

MDP solution: Terms

- Execution history
- History-dependent policy
- Markovian policy
- Stationary Markovian policy
- Value function
- Optimal MDP solution

Policy

We need a policy, a rule for action selection that works in any state

Search problem: path (sequence of actions)

MDP:



Definition: policy

A **policy** π is a mapping from each state $s \in \text{States}$ to an action $a \in \text{Actions}(s)$.

What information might such a rule use?

Definition 2.2 Execution History. An *execution history* of an MDP up to time step $t \in \mathcal{D}$ is a sequence $h_t = ((s_1, a_1), \dots, (s_{t-1}, a_{t-1}), s_t)$ of pairs of states the agent has visited and actions the agent has chosen in those states for all time steps t' s.t. $1 \leq t' \leq t - 1$, plus the state visited at time step t .

History-Dependent and Markovian Policy

Definition 2.3 History-Dependent Policy. A *probabilistic history-dependent policy* for an MDP is a probability distribution $\pi : \mathcal{H} \times \mathcal{A} \rightarrow [0, 1]$ that assigns to action $a \in \mathcal{A}$ a probability $\pi(h, a)$ of choosing it for execution at the current time step if the execution history up to the current time step is $h \in \mathcal{H}$. A *deterministic history-dependent policy* is a mapping $\pi : \mathcal{H} \rightarrow \mathcal{A}$ that assigns to each $h \in \mathcal{H}$ an action $a \in \mathcal{A}$ to be executed at the current time step if h the execution history up to the current time step.

A history-dep policy must provide a distb over actions for every possible history

Representing history-dependent policies

Definition 2.4 Markovian Policy. A probabilistic (deterministic) history-dependent policy $\pi : \mathcal{H} \times \mathcal{A} \rightarrow [0, 1]$ ($\pi : \mathcal{H} \rightarrow \mathcal{A}$) is *Markovian* if for any two histories $h_{s,t}$ and $h'_{s,t}$, both of which end in the same state s at the same time step t , and for any action a , $\pi(h_{s,t}, a) = \pi(h'_{s,t}, a)$ ($\pi(h_{s,t}) = a$ if and only if $\pi(h'_{s,t}) = a$).

For several important types of MDPs and objectives, at least one optimal soln is necessarily Markovian

Stationary Markovian Policy

If $\mathcal{D}$ is finite, size of policy specification is at most linear in $|\mathcal{D}|$

Definition 2.5 Stationary Markovian Policy. A probabilistic (deterministic) Markovian policy $\pi : \mathcal{S} \times \mathcal{D} \times \mathcal{A} \rightarrow [0, 1]$ ($\pi : \mathcal{S} \times \mathcal{D} \rightarrow \mathcal{A}$) is *stationary* if for any state s , action a , and two time steps t_1 and t_2 , $\pi(s, t_1, a) = \pi(s, t_2, a)$ ($\pi(s, t_1) = a$ if and only if $\pi(s, t_2) = a$), i.e., π does not depend on time.

$$\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1] \quad \pi : \mathcal{S} \rightarrow \mathcal{A}$$

Practically interesting MDPs with infinite number of time steps, at least one optimal soln that is stationary

Definition 2.6 Value Function. A *history-dependent value function* is a mapping $V : \mathcal{H} \rightarrow [-\infty, \infty]$. A *Markovian value function* is a mapping $V : \mathcal{S} \times \mathcal{D} \rightarrow [-\infty, \infty]$. A *stationary Markovian value function* is a mapping $V : \mathcal{S} \rightarrow [-\infty, \infty]$.

Value function and Optimality

Definition 2.7 The Value Function of a Policy. Let $h_{s,t}$ be a history that terminates at state s and time t . Let $R_{t'}^{\pi_{h_{s,t}}}$ be random variables for the amount of reward obtained in an MDP as a result of executing policy π starting in state s for all time steps $t' \text{ s.t. } t \leq t' \leq |\mathcal{D}|$ if the MDP ended up in state s at time t via history $h_{s,t}$. The value function $V^\pi : \mathcal{H} \rightarrow [-\infty, \infty]$ of a history-dependent policy π is a utility function u of the reward sequence $R_t^{\pi_{h_{s,t}}}, R_{t+1}^{\pi_{h_{s,t}}}, \dots$ that one can accumulate by executing π at time steps $t, t+1, \dots$ after history $h_{s,t}$. Mathematically, $V^\pi(h_{s,t}) = u(R_t^{\pi_{h_{s,t}}}, R_{t+1}^{\pi_{h_{s,t}}}, \dots)$.

Exact way of converting a seq of rewards into utility yet to be made concrete

$$V^\pi(s, t) = u(R_t^{\pi_{s,t}}, R_{t+1}^{\pi_{s,t}}, \dots) \quad V^\pi(s) = u(R_t^{\pi_s}, R_{t+1}^{\pi_s}, \dots)$$

Definition 2.8 Optimal MDP Solution. An *optimal solution* to an MDP is a policy π^* s.t. the value function of π^* , denoted as V^* and called *the optimal value function*, dominates the value functions of all other policies for all histories h . Mathematically, for all $h \in \mathcal{H}$ and any π , V^* must satisfy $V^*(h) \geq V^\pi(h)$.

Caveats: a) Utility unspecified, b) Comparing value functions

$$u(R_t, R_{t+1}, \dots) = \sum_{t'=t}^{|\mathcal{D}|} R_{t'}$$

Expected Linear Additive Utility

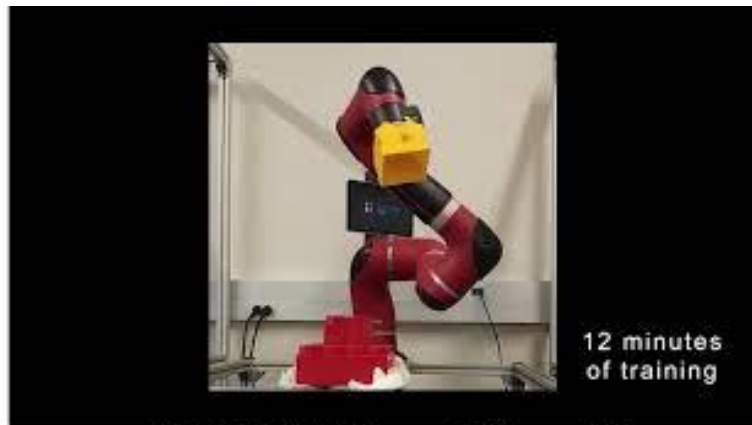
For an arbitrary u , the optimal value function need not exist

Definition 2.9 Expected Linear Additive Utility. An *expected linear additive utility* function is a function $u(R_t, R_{t+1}, \dots) = \mathbb{E}[\sum_{t'=t}^{|\mathcal{D}|} \gamma^{t'-t} R_{t'}] = \mathbb{E}[\sum_{t'=0}^{|\mathcal{D}|-t} \gamma^{t'} R_{t'+t}]$ that computes the utility of a reward sequence as the expected sum of (possibly discounted) rewards in this sequence, where $\gamma \geq 0$ is the *discount factor*.

Caveat: Assumes agent to be risk-neutral

Discounting

Episodic



© Authors of ICRA 2018 Paper 1789

Thu AM

Post Q.2

[Source: Youtube](#)

Discount factor: $\gamma \in [0, 1]$

Infinite horizon



[Source: Youtube](#)

Infinite horizon: $\gamma \in [0, 1)$

Solution Existence

***The Optimality Principle.** If every policy's quality can be measured by this policy's expected linear additive utility, there exists a policy that is optimal at every time step.*

Definition 2.10 Finite-Horizon MDP. A finite-horizon MDP is an MDP as described in Definition 2.1 with a finite number of time steps, i.e., with $|\mathcal{D}| = T_{max} < \infty$.

Real-world scenario: Teacher trying to teach a set of skills to a student over T_{max} lessons

Optimality Principle: FH MDP

Theorem 2.11 The Optimality Principle for Finite-Horizon MDPs. For a finite-horizon MDP with $|\mathcal{D}| = T_{\max} < \infty$, define $V^\pi(h_{s,t}) = \mathbb{E}[\sum_{t'=0}^{T_{\max}-t} R_{t'+t}^{\pi_{h_{s,t}}}]$ for all $1 \leq t \leq T_{\max}$, and $V^\pi(h_{s,T_{\max}+1}) = 0$. Then the optimal value function V^* for this MDP exists, is Markovian, and satisfies, for all $h \in \mathcal{H}$,

$$V^*(h) = \max_{\pi} V^\pi(h)$$

and, for all $s \in \mathcal{S}$ and $1 \leq t \leq T_{\max}$,

$$V^*(s, t) = \max_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s', t) [\mathcal{R}(s, a, s', t) + V^*(s', t+1)] \right]. \quad (2.2)$$

Moreover, the optimal policy π^* corresponding to the optimal value function is deterministic Markovian and satisfies, for all $s \in \mathcal{S}$ and $1 \leq t \leq T_{\max}$,

$$\pi^*(s, t) = \operatorname{argmax}_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s', t) [\mathcal{R}(s, a, s', t) + V^*(s', t+1)] \right]. \quad (2.3)$$

Deterministic, Markovian optimal policies for finite-horizon MDPs

Optimal value function dominates all other value functions, including general history-dependent ones

Infinite Horizon

Definition 2.12 Stationary Transition Function. An MDP's transition function $\mathcal{T}$ is stationary if for any states $s_1, s_2 \in \mathcal{S}$ and action $a \in \mathcal{A}$, the value $\mathcal{T}(s_1, a, s_2, t)$ does not depend on t , i.e., $\mathcal{T}(s_1, a, s_2, t) = \mathcal{T}(s_1, a, s_2, t')$ for any $t, t' \in \mathcal{D}$.

Definition 2.13 Stationary Reward Function. An MDP's reward function $\mathcal{R}$ is stationary if for any states $s_1, s_2 \in \mathcal{S}$ and action $a \in \mathcal{A}$, the value $\mathcal{R}(s_1, a, s_2, t)$ does not depend on t , i.e., $\mathcal{R}(s_1, a, s_2, t) = \mathcal{R}(s_1, a, s_2, t')$ for any $t, t' \in \mathcal{D}$.

$$\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1] \text{ and } \mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$$

set γ to a value $0 \leq \gamma < 1$

$$u(R_t, R_{t+1}, \dots) \leq \sum_{t'=0}^{\infty} \gamma^{t'} M = \frac{M}{1-\gamma}$$

Infinite Horizon

Definition 2.14 Infinite-Horizon Discounted-Reward MDP. An infinite-horizon discounted-reward MDP is a stationary MDP as described in Definition 2.1, in which the value of a policy is defined as $V^\pi(h_{s,t}) = \mathbb{E}[\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'+t}^{\pi_{h_{s,t}}}]$, where the discount factor γ is a model parameter restricted to be $0 \leq \gamma < 1$.

Infinite Horizon MDP: Optimality Principle

Theorem 2.15 The Optimality Principle for Infinite-Horizon MDPs. For an infinite-horizon discounted-reward MDP with discount factor γ s.t. $0 \leq \gamma < 1$, define $V^\pi(h_{s,t}) = \mathbb{E}[\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'+t}^{\pi_{h_{s,t}}}]$. Then the optimal value function V^* for this MDP exists, is stationary Markovian, and satisfies, for all $h \in \mathcal{H}$,

$$V^*(h) = \max_{\pi} V^\pi(h) \quad (2.4)$$

and, for all $s \in \mathcal{S}$,

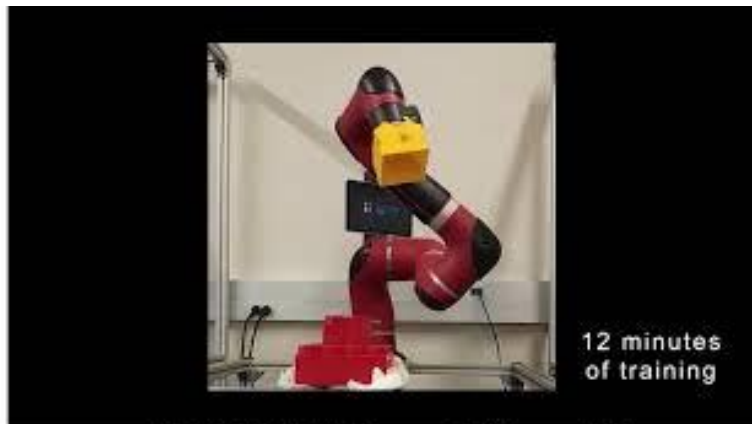
Optimal policy is Deterministic, Stationary, Markovian

$$V^*(s) = \max_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') [\mathcal{R}(s, a, s') + \gamma V^*(s')] \right]. \quad (2.5)$$

Moreover, the optimal policy π^* corresponding to the optimal value function is deterministic stationary Markovian and satisfies, for all $s \in \mathcal{S}$,

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') [\mathcal{R}(s, a, s') + \gamma V^*(s')] \right]. \quad (2.6)$$

Indefinite Horizon MDP



© Authors of ICRA 2018 Paper 1795

The AM

Pod Q.2

[Source](#)



[Source](#)

Total expected cost of any optimal policy in an indefinite horizon MDP should be finite even without a discount factor

Indef H MDPs: Proper Policy

Definition 2.16 Proper policy. For a given stationary MDP, let $\mathcal{G} \subseteq \mathcal{S}$ be a set of *goal* states s.t. for each $s_g \in \mathcal{G}$ and all $a \in \mathcal{A}$, $\mathcal{T}(s_g, a, s_g) = 1$ and $\mathcal{R}(s_g, a, s_g) = 0$. Let h_s be an execution history that terminates at state s . For a given set $S' \subseteq \mathcal{S}$, let $P_t^\pi(h_s, S')$ be the probability that after execution history h_s , the agent transitions to some state in S' within t time steps if it follows policy π . A policy π is called *proper at state s* if $P_t^\pi(h_s, \mathcal{G}) = 1$ for all histories $h_s \in \mathcal{H}$ that terminate at s .

If for some h_s , $\lim_{t \rightarrow \infty} P_t^\pi(h_s, \mathcal{G}) < 1$, π is called *improper at state s* . A policy π is called *proper* if it is proper at all states $s \in \mathcal{S}$. Otherwise, it is called *improper*.

A proper policy is indefinite-horizon

Stochastic Shortest Path MDPs

Definition 2.17 Stochastic Shortest-Path MDP. (*Weak definition.*) A *stochastic shortest-path (SSP) MDP* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{C}, \mathcal{G} \rangle$ where:

- $\mathcal{S}$ is the finite set of all possible states of the system,
- $\mathcal{A}$ is the finite set of all actions an agent can take,
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a stationary transition function specifying the probability $\mathcal{T}(s_1, a, s_2)$ of going to state s_2 whenever action a is executed when the system is in state s_1 ,
- $\mathcal{C} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, \infty)$ is a stationary cost function that gives a finite strictly positive cost $\mathcal{C}(s_1, a, s_2) > 0$ incurred whenever the system goes from state s_1 to state s_2 as a result of executing action a , with the exception of transitions from a goal state, and
- $\mathcal{G} \subseteq \mathcal{S}$ is the set of all goal states s.t. for every $s_g \in \mathcal{G}$, for all $a \in \mathcal{A}$, and for all $s' \notin \mathcal{G}$, the transition and cost functions obey $\mathcal{T}(s_g, a, s_g) = 1$, $\mathcal{T}(s_g, a, s') = 0$, and $\mathcal{C}(s_g, a, s_g) = 0$,

under the following condition:

- There exists at least one proper policy.

SSP MDPs: Optimality Principle

Theorem 2.18 The Optimality Principle for Stochastic Shortest-Path MDPs. For an SSP MDP, define $V^\pi(h_{s,t}) = \mathbb{E}[\sum_{t'=0}^{\infty} R_{t'+t}^{\pi_{h_{s,t}}}]$. Then the optimal value function V^* for this MDP exists, is stationary Markovian, and satisfies, for all $h \in \mathcal{H}$,

$$V^*(h) = \min_{\pi} V^\pi(h) \quad (2.7)$$

and, for all $s \in \mathcal{S}$,

Optimal policy is just a mapping from states to actions

$$V^*(s) = \min_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') [\mathcal{C}(s, a, s') + V^*(s')] \right]. \quad (2.8)$$

Moreover, the optimal policy π^* corresponding to the optimal value function is deterministic stationary Markovian and satisfies, for all $s \in \mathcal{S}$,

How many deterministic stationary Markovian policies are there?

$$\pi^*(s) = \operatorname{argmin}_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') [\mathcal{C}(s, a, s') + V^*(s')] \right]. \quad (2.9)$$

SSP MDPs: Optimal Policy

Theorem 2.19 In an SSP MDP, $V^*(s)$ is the smallest expected cost of getting from state s to the goal. Every optimal stationary deterministic Markovian policy in an SSP MDP is proper and minimizes the expected cost of getting to the goal at every state.

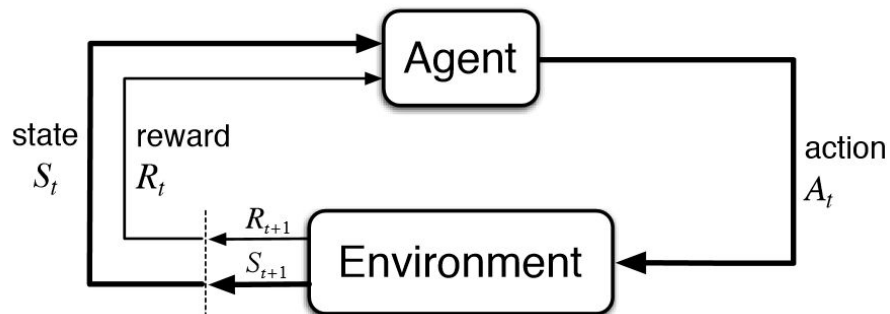
Min expected linear additive utility leads to min expected cost from state s to goal

SSP MDPs and other MDPs

Theorem 2.21 Denote the set of all finite-horizon MDPs as FH , the set of all infinite-horizon discounted-reward MDPs as IFH , and the set of all SSP MDPs as SSP . The following statements hold true [24]:

$$FH \subset SSP$$

$$IFH \subset SSP$$



Value Functions (& notation)

Value Functions

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t \mid S_t=s] = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t=s\right], \text{ for all } s \in \mathcal{S}, \quad (3.12)$$

$$q_{\pi}(s, a) \doteq \mathbb{E}_{\pi}[G_t \mid S_t=s, A_t=a] = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t=s, A_t=a\right]. \quad (3.13)$$