Theory of Average-Reward Markov Decision Processes

Mohammad Sadegh Talebi
m.shahi@di.ku.dk
Department of Computer Science

Average-Reward MDPs

Recall the definition of a generic MDP model: $M = \big(\mathcal{S}, \mathcal{A}, P, R\big)$

- State-space $\mathcal{S}$
- Action-space $\mathcal{A} = \cup_{s\in\mathcal{S}}\mathcal{A}_s$
  - $\mathcal{A}_s$ is the set of actions available in state s
- Transition function P: Selecting $a \in \mathcal{A}_s$ in $s \in \mathcal{S}$ leads to a transition to s' with probability $P(s'|s,a)$. $P(\cdot|s,a)$ is a probability distribution over $\mathcal{S}$, i.e.,

$$\sum_{s'} P(s'|s,a) = 1$$

- Reward function R: Selecting $a \in \mathcal{A}_s$ in $s \in \mathcal{S}$ yields a reward $r \sim R(s,a)$.

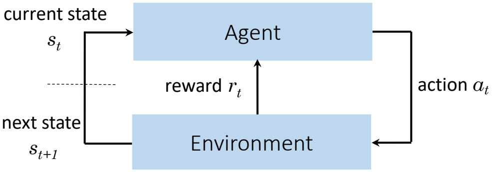
For simplicity, we consider an identical action set all states, i.e., $\mathcal{A}_s = \mathcal{A}$ for all $s \in \mathcal{S}$.

Interaction with MDP

An **agent** interacts with the MDP for N rounds.

At each time step t:

- The agent observes the current state s_t and takes an action $a_t \in \mathcal{A}_{s_t}$
- The environment (MDP) decides a reward $r_t := r(s_t, a_t) \sim R(s_t, a_t)$ and a next state $s_{t+1} \sim P(\cdot | s_t, a_t)$
- The agent receives r_t (any time in step t before start of $t+1$)



This interaction produces a trajectory (or history)

$$h_t = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t)$$

Classification of MDPs based on N

- **Finite-Horizon MDPs**: $N < \infty$, and the goal is to solve

$$\max_{\text{all strategies}} \mathbb{E}\left[\sum_{t=1}^{N-1} r(s_t, a_t) + r(s_N)\right]$$

- **Infinite-Horizon Discounted MDPs:** $N = \infty$, and given discount factor $\gamma \in (0,1)$, the goal is to solve

$$\max_{\text{all strategies}} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)\right]$$

- **Infinite-Horizon Undiscounted MDPs (Average-Reward MDPs):** $N = \infty$, and the goal is to solve

$$\max_{\text{all strategies}} \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{t=1}^{N} r(s_t, a_t)\right]$$

This lecture: We study Average-Reward MDPs.

The Optimality Criterion

Let's consider optimizing the N-step cumulative reward (a.k.a. total reward):

$$\sup_{\text{all strategies}} \mathbb{E}\left[\sum_{t=1}^{N} r_t\right]$$

$\Longrightarrow$ An ill-defined objective as it could grow unbounded when $N \to \infty$, *even with bounded rewards*.

We instead consider maximizing the average expected reward:

$$\sup_{\text{all strategies}} \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{t=1}^{N} r_t\right]$$

- Hence the name average-reward MDPs.
- A well-defined objective.
- It also makes sense in practice. (More on this later.)

Assumption on Rewards

We assume:

- Deterministic rewards so that

$$r_t = r(s_t, a_t) = R(s_t, a_t)$$

- Bounded rewards in the following sense:

$$|r(s,a)| \le R_{\max} < \infty$$

Extension to stochastic reward is done by replacing $r(s,a)$ with the $\mathbb{E}[r(s,a)]$ in the results.

Policy

When interacting with an MDP, actions are taken according to some policy. Policies classes are defined identically as in discounted MDPs, where a policy may be:

- deterministic or randmozied (stochastic)
- history-dependent or stationary

	Deterministic	Randomized
Stationary	$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{S} \to \Delta(\mathcal{A})$
History-dependent	$\pi : \mathcal{H}_t \to \mathcal{A}$	$\pi : \mathcal{H}_t \to \Delta(\mathcal{A})$

- $\Delta(\mathcal{A})$ denotes the simplex of probability distributions over $\mathcal{A}$.
- $\mathcal{H}_t$ the set of all possible history sequences up to time t
- For a randomized policy π, $\pi(a|s)$ denotes the probability of choosing a in s.

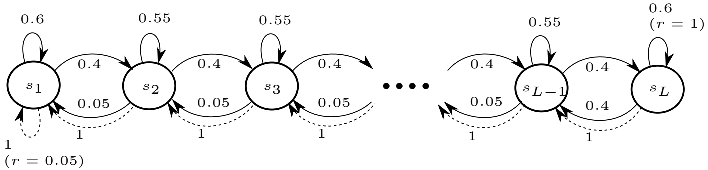
Policy

	Deterministic	Randomized
Stationary	$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{S} \to \Delta(\mathcal{A})$
History-dependent	$\pi : \mathcal{H}_t \to \mathcal{A}$	$\pi : \mathcal{H}_t \to \Delta(\mathcal{A})$

- Π^{SD}: Stationary deterministic policies
- Π^{SR}: Stationary randomized policies
- Π^{HD}: History-dependent deterministic policies
- Π^{HR}: History-dependent randomized policies

$$\text{(i) } \Pi^{\mathsf{SD}} \subset \Pi^{\mathsf{SR}} \subset \Pi^{\mathsf{HR}}$$
$$\text{(ii) } \Pi^{\mathsf{SD}} \subset \Pi^{\mathsf{HD}} \subset \Pi^{\mathsf{HR}}$$

Example 1

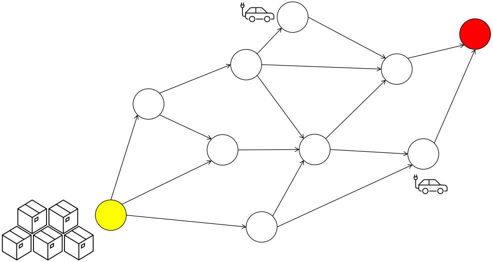


A continual task in RiverSwim

- **Variant 1:** The agent interacts with RiverSwim for an unspecified number N of round.
- **Variant 2:** If in s_L and taking 'right', *Kystvagten* brings the agent to a random state, and the task repeats —the corresponding transition is not shown here.

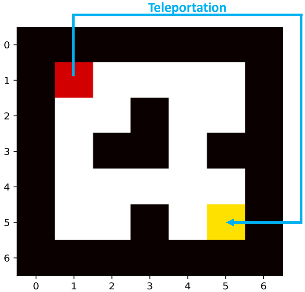
Can you guess an optimal policy in either variants?

Example 2



- Task: Transporting an arbitrary number of packages between source (in yellow) and destination (in red).
- The transportation cost differs across paths, and we are interested in minimizing the total cost.
- One package per round. Occasionally, a charging station must be visited.

Example 3



- A grid-world with $S = 20$ states, and 4 actions (Up, Down, Left, Right).
- E.g., 'Up' yields: moving up (w.p. 0.7), no move (w.p. 0.1), or moving left or right (each w.p. 0.1) —walls act as reflector.
- Reward is zero everywhere, except in the goal state (in red).
- The task is **continual**: Once in the goal state, the agent is teleported to the initial state.

Gain and Bias

Gain vs. Value

- For discounted and finite-horizon MDPs we defined notions of value function to distinguish the quality of various policies.
- Value functions measure the sum of future (discounted) rewards starting from any state.
- This machinery does not carry over to average-reward MDPs as cumulative reward could grow without bound.
- Instead, we define the notion of gain and bias to rank policies.

Gain

The gain function of policy π is a mapping $g^\pi : \mathcal{S} \to \mathbb{R}$ defined as

$$g^\pi(s) := \lim_{N\to\infty} \frac{1}{N}\mathbb{E}^\pi\left[\sum_{t=1}^{N} r(s_t, a_t)\middle| s_1 = s\right].$$

where $\mathbb{E}^\pi$ indicates expectation over trajectories generated by π.

- $g^\pi(s)$ measure the per-step reward obtained under π starting from s, in the long run.
- The limit may not exist for all policies.
- For all π and s:

$$|g^\pi(s)| \leq R_{\max} ,$$

where $R_{\max}$ is an upper bound on the rewards.

Optimization using Gains

Solving an average-reward MDP M amounts to solving the following optimization problem:

$$g^\star(s) = \sup_{\pi \in \Pi^{\mathsf{HR}}} g^\pi(s)\,,$$

for all $s \in \mathcal{S}$.

- $g^\star : \mathcal{S} \to \mathbb{R}$ is called the optimal gain.
- Any policy achieving $g^\star(s)$ for all s is called gain-optimal (or optimal, for short) and denoted by $\pi^\star$.
- Do we have other optimality criteria? Discussion in class.

Is Gain Sufficient?

- Is gain alone is sufficient? $\Longrightarrow$ Yes, if only the steady-state regime of MDP is concerned.
- However, for finite N,

$$\mathbb{E}^\pi\left[\sum_{t=1}^{N} r(s_t, a_t)\middle| s_1 = s\right] \quad \neq \quad Ng^\pi(s)$$

- The difference $\mathbb{E}^\pi\left[\sum_{t=1}^{N} r(s_t, a_t)\middle| s_1 = s\right] - Ng^\pi(s)$ reflects the transient rewards.
- To capture the difference due to the transient regime we define bias.

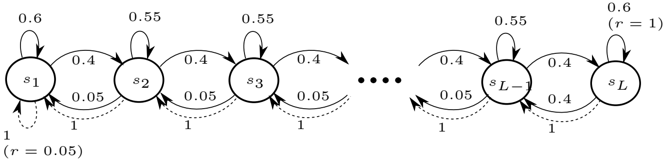
Bias

The bias function (or simply, bias) of policy π is a mapping $b^\pi : \mathcal{S} \to \mathbb{R}^S$ defined as

$$b^\pi(s) := \mathbb{E}^\pi\left[\sum_{t=1}^{\infty}\Big(r(s_t,a_t) - g^\pi(s_1)\Big)\Big|s_1 = s\right].$$

where $\mathbb{E}^\pi$ indicates expectation over trajectories generated by π.

- Assume $g^\pi(s) = g$ is constant, i.e., the MDP *forgets* the initial state —for example, it holds in RiverSwim for π prescribing to take 'right' action.
- Then $b^\pi(s) - b^\pi(s')$ indicates how much reward could have been obtained by starting in s rather than in s'.



Digression: Classification of MDPs Based on Reachability

MDP Classes

A classification of MDPs in terms of reachability of various states:

1. An MDP is **ergodic** if it is possible to reach any state from any other state under every $\pi \in \Pi^{\mathsf{SD}}$.

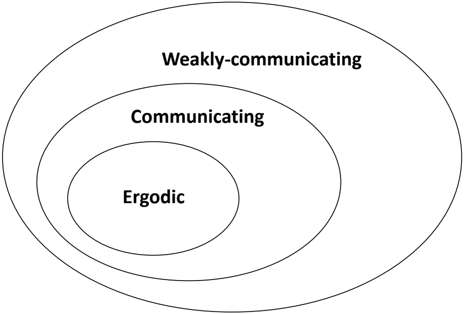
2. An MDP is **communicating** if it is possible to reach any state from any other state under some $\pi \in \Pi^{\mathsf{SD}}$.

3. An MDP is **weakly communicating** if its state-space can be partitioned into two sets:
   (i) a set that is transient under every $\pi \in \Pi^{\mathsf{SD}}$; and
   (ii) a closed set in which every two states can reach each other under some $\pi \in \Pi^{\mathsf{SD}}$.

   In words, a weakly communicating MDP $\equiv$ a communicating MDP + some extra transient states.

MDP Classes

Hierarchy of MDP classes:



Diameter

Connectivity in MDPs can be measured via diameter (Jaksch et al., 2010).

Diameter of MDP

Let $T^\pi(s', s)$ denote the first hitting time of state s' when following $\pi \in \Pi^{\mathsf{SD}}$ from $s (\neq s')$ in an MDP M. The diameter D of M is defined as

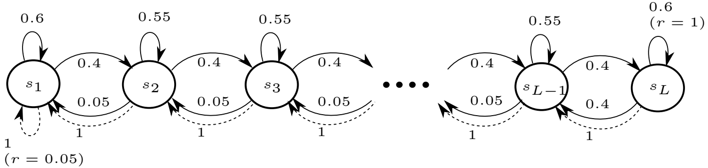
$$D := \max_{s \neq s'} \min_{\pi \in \Pi^{\mathsf{SD}}} \mathbb{E}\big[T^\pi(s', s)\big].$$

- Intuitively, D measures the worst-case shortest-path in the MDP:

$$D := \underbrace{\max_{s \neq s'}}_{\text{worst-case}} \underbrace{\min_{\pi \in \Pi^{\mathsf{SD}}} \mathbb{E}\big[T^\pi(s', s)\big]}_{\text{shortest-path for } s \to s'}.$$

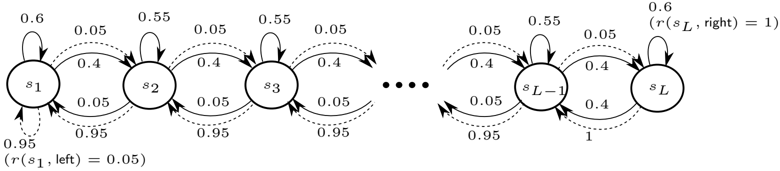
- MDP M is communicating $\iff$ M has a finite diameter.
- We may have $D = \infty$ for a weakly communicating MDP.

Example: RiverSwim



Is this MDP ergodic? Is it communicating?

Example: Ergodic RiverSwim



Is this MDP ergodic? Is it communicating?

MDP Classes: Gain

- In ergodic MDPs: g^π for any π does not depend on the starting state, i.e., $g^\pi(s) = g^\pi$ for all s.
- In weakly-communicating MDPs: $g^\star$ does not depend on the starting state, i.e., $g^\star(s) = g^\star$ for all s.

	ergodic	communicating	weakly-communicating
g^π	constant	(maybe) state-dependent	(maybe) state-dependent
$g^\star$	constant	constant	constant
D	finite	finite	(maybe) infinite

From now on, we only consider weakly-communicating MDPs.

Finding a Gain-Optimal Policy

Optimal Policy

In a weakly-communicating MDPs, at least one stationary deterministic policy exists, which is gain-optimal.

Hence,

$$g^\star(s) = g^\star = \sup_{\pi \in \Pi^{\mathsf{HR}}} g^\pi(s) = \max_{\pi \in \Pi^{\mathsf{SD}}} g^\pi(s)$$

- Hence, we can restrict attention to $\pi \in \Pi^{\mathsf{SD}}$.
- Such optimal policy in $\pi \in \Pi^{\mathsf{SD}}$ can be characterized using Bellman optimality equations.

Bellman Optimality Equations

Theorem

If M is weakly communicating, then:

$$g^\star + b^\star(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \sum_{x \in \mathcal{S}} P(x|s,a) b^\star(x) \right), \quad \forall s \in \mathcal{S}.$$

Furthermore, $\pi \in \Pi^{SD}$ is optimal if and only if:

$$\pi(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left(r(s,a) + \sum_{x \in \mathcal{S}} P(x|s,a) b^\star(x) \right), \quad \forall s \in \mathcal{S}.$$

- $b^\star : \mathcal{S} \to \mathbb{R}$ is called the optimal bias function.
- $g^\star$ is uniquely defined. (Why?)
- But $b^\star$ is defined up to an additive constant: If $b^\star$ is a solution, so is $b^\star + c\mathbf{1}$ for any $c \in \mathbb{R}$.

VI

- We can use Value Iteration to solve Bellman Optimality Equations.
- The update is similar to the one in discounted MDPs:

$$V_{n+1}(s) = \max_{a\in\mathcal{A}} \Big(r(s,a) + \sum_{x\in\mathcal{S}} P(x|s,a)V_n(x) \Big), \quad s \in \mathcal{S}.$$

- V_n could grow unbounded. Yet we can show that $V_{n+1} - V_n$ could converge (to $g^\star$).
- Hence, we choose to stop as soon as

$$\max_{s\in\mathcal{S}} \big(V_{n+1}(s) - V_n(s)\big) - \min_{s\in\mathcal{S}} \big(V_{n+1}(s) - V_n(s)\big) < \varepsilon$$

Or $\mathrm{sp}(V_{n+1} - V_n) < \varepsilon$, where 'sp' denotes the span operator (or span semi-norm) defined as

$$\text{Given } f : \mathcal{S} \to \mathbb{R}^S, \qquad \mathrm{sp}(f) := \max_{s\in\mathcal{S}} f(s) - \min_{s\in\mathcal{S}} f(s)\,.$$

VI

- **input:** ε
- **initialization:** Select $V_0 \in \mathbb{R}^S$ arbitrarily. Set $n = -1$.
- **repeat:**
  - Increment n
  - Update, for each $s \in \mathcal{S}$,

$$V_{n+1}(s) = \max_{a \in \mathcal{A}} \Big(r(s,a) + \sum_{x \in \mathcal{S}} P(x|s,a) V_n(x) \Big)$$

  until $\mathrm{sp}\big(V_{n+1} - V_n\big) < \varepsilon$
- **output:**

$$\pi^{\mathtt{VI}}(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \Big(r(s,a) + \sum_{x \in \mathcal{S}} P(x|s,a) V_n(x) \Big), \quad s \in \mathcal{S}$$

VI: Convergence

Theorem

In weakly communicating MDPs,

- *For any $V_0 \in \mathbb{R}$, $(V_n)_{n\geq 0}$ generated by* VI *satisfies,*

$$\lim_{n\to\infty} \big(V_{n+1}(s) - V_n(s)\big) = g^\star\,, \quad \forall s \in \mathcal{S}.$$

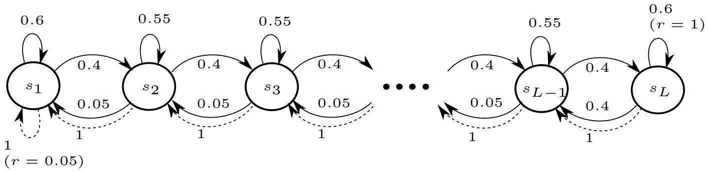
- VI *converges after finitely many iterations. Furthermore, π^{VI} is ε-optimal: For all $s \in \mathcal{S}$, $g^{\pi^{VI}}(s) \geq g^\star - \varepsilon$.*

- $V_{n+1}(s) - V_n(s)$ for any s gives an approximation to $g^\star$. It is best to approximate $g^\star$ as:

$$\frac{1}{2}\Big[\max_{s\in\mathcal{S}}(V_{n+1}(s) - V_n(s)) + \min_{s\in\mathcal{S}}(V_{n+1}(s) - V_n(s))\Big]$$

- V_n also gives an approximation for $b^\star$. So does $V_n - (\min_s V_n(s))\mathbf{1}$ (why?)

Example: RiverSwim



Optimal gain and optimal bias function in 6-state RiverSwim, computed via `VI`:

$$g^\star = 0.467$$

$$b^\star(s_1) = 0, \quad b^\star(s_2) = 0.78, \quad b^\star(s_3) = 2.04$$

$$b^\star(s_4) = 3.37, \quad b^\star(s_5) = 4.70, \quad b^\star(s_6) = 6.03$$

Total Reward and Gain

N-step total reward, $\sum_{t=1}^{N} r_t$ is naturally connected to the average-reward.

Theorem

In weakly communicating MDPs, under $\pi^\star$,

$$(i) \quad \mathbb{E}\left[\sum_{t=1}^{N} r_t \middle| s_1 = s\right] = Ng^\star + \mathcal{O}\big(\mathrm{sp}(b^\star)\big)$$

$$(ii) \quad \sum_{t=1}^{N} r_t = Ng^\star + \mathcal{O}\Big(\mathrm{sp}(b^\star)\sqrt{N\log(N/\delta)}\Big), \quad \textit{w.p.} \ \geq 1-\delta$$

(i) is evident from the definition of bias function, and (ii) follows from Hoeffding's inequality.

Evaluating Gain and Bias

(outside of the scope of OReL)

Induced MRPs

Every $\pi \in \Pi^{\mathsf{SR}}$ induces a Markov reward process (MRP) —defined identically as in discounted MDPs.

- The transition matrix P^π of MRP:

$$P^\pi(s,s') = \sum_{a\in\mathcal{A}} \pi(a|s)P(s'|s,a)\,, \quad s,s' \in \mathcal{S}$$

- The reward vector of r^π of MRP:

$$r^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)r(s,a)\,, \quad s \in \mathcal{S}$$

Gain of Stationary Policies

Theorem

Let $\pi \in \Pi^{SR}$. *Then,* $g^\pi = \overline{P}^\pi r^\pi$, *where*

$$\overline{P}^\pi := \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} (P^\pi)^{t-1}$$

is the limiting matrix or the Cesaro-average of P^π,

Proof. For $N \in \mathbb{N}$, the N-step accumulated reward in the MRP induced by π is

$$\mathbb{E}^\pi \Big[\sum_{t=1}^{N} r(s_t, a_t) \Big| s_1 = s \Big]$$

$$= \mathbb{E}^\pi [r(s_1, a_1)|s_1 = s] + \mathbb{E}^\pi [r(s_2, a_2)|s_1 = s] + \ldots + \mathbb{E}^\pi [r(s_N, a_N)|s_1 = s\Big]$$

$$= r^\pi(s) + [P^\pi r^\pi](s) + \ldots + [(P^\pi)^{N-1} r^\pi](s) = \sum_{t=1}^{N} \big[(P^\pi)^{t-1} r^\pi \big](s)$$

where we used that for any $t \geq 1$, when following $\pi \in \Pi^{\mathrm{SR}}$,

$$\mathbb{P}(s_t = y | s_1 = x) = (P^\pi)^t(x, y).$$

Bellman Equation for Π^{SR}

Theorem (Bellman Equation for Policy π)

Let $\pi \in \Pi^{SR}$. Assume that π induces an MRP, which is irreducible or unichain. Then

$$g^\pi \mathbf{1} = \overline{P}^\pi r^\pi$$

Furthermore, the bias function b^π satisfies the Bellman equation:

$$g^\pi \mathbf{1} + (I - P^\pi) b^\pi = r^\pi .$$

As a result,

$$b^\pi = (I - P^\pi + \overline{P}^\pi)^{-1}(I - \overline{P}^\pi) r^\pi + c\mathbf{1},$$

where c is any arbitrary scalar.

Note that the matrix $I - P^\pi + \overline{P}^\pi$ is non-singular, so the last assertion above is well-defined.