Markov chains

The uncertainty about the complex dynamics of the environment is modelled via Markov processes.

↳ Simplifying assumption: Markov property → the future dynamics of the environment is independent of the past given the present state. Formally, denoting by S_t the state of the environment at time t, we have:

$$\mathbb{P}[S_{t+1} \mid S_t, S_{t-1}, ..] = \mathbb{P}[S_{t+1} \mid S_t]$$

Here S will be assumed to belong to a finite state space $\mathcal{S}$.

↳ Transition probabilities: $P : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ such that

* $P_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$: probability to transition from state s to successor state s'.

* In general, transition can be time-dependent, but will be assumed stationary here.

* Concretely $P_{ss'}$ defines a stochastic matrix, ie a matrix with non negative real entries which sum to one row-wise:

$$\underline{\underline{P}} = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & & \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \text{ with } \left. \begin{array}{l} \sum_{s'} P_{ss'} = 1 \\ P_{ss'} \geqslant 0 . \end{array} \right) \text{ model for the environment}$$

↳ Stationary dynamics: Under well-known condition (irreducibility) Markovian dynamics are ergodic, i.e., the probability distribution of S_t converges toward a unique distribution π_∞ independently from the starting distribution π_0. In other words:

$$\pi_t = \pi_0 P^t = \pi_0 (P \circ P \circ \ldots \circ P) \xrightarrow[t \to +\infty]{} \pi_\infty \quad \text{at least weakly}$$

The invariant measure may be concentrated on one (absorbing) state. In this case, interest bears on the "transients" of the environment dynamics

Markov reward process

Consider a Markov chain $(\mathcal{S}, P)$ with finite state space $\mathcal{S}$ and transition matrix P.

A Markov reward process is given by $(\mathcal{S}, P, R, \gamma)$ where

1) R is a *reward function*, i.e., a random variable that follows the Markov property:

$$\mathbb{E}[R_{t+1} \mid S_t, S_{t-1}, \dots] = \mathbb{E}[R_{t+1} \mid S_t] = \mathcal{R}_{S_t} \leftarrow \text{state dependent expected reward}$$

2) γ is a *discount factor* $0 \leq \gamma \leq 1$ which measures the preference for immediate reward and is introduced to controlled the aggregated future reward / return:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+2} + \dots = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1}$$

Markov reward setting

Here, infinite horizon with discount rate: the performance is measured via an infinite series aggregated future reward discounted geometrically. Other settings are possible: for instance finite horizon without discounting, infinite horizon with average bounded reward, random horizon with terminal state ....

Value Function

The value function of a Markov reward process measures the expected future return given the current state: $v : \mathcal{S} \to \mathbb{R}$

$$v(s) = \mathbb{E}[G_t \mid S_t = s] = \mathbb{E}\Big[\sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} \;\Big|\; S_t = s\Big]$$

(No concept of optimization at this stage.)

Bellman equation for Markov reward process

Goal: given a MRP $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$, compute the value function $v(s)$ which measure how "beneficial" a state is.

Bellman idea: value function = immediate reward + discounted value at successor state

$$v(s) = \mathbb{E}\left[G_t \mid S_t = s\right]$$

$$= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right]$$

$$= \mathbb{E}\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \dots\right) \mid S_t = s\right]$$

$$= \mathbb{E}\left[R_{t+1} \mid S_t = s\right] + \gamma\, \mathbb{E}\left[G_{t+1} \mid S_t = s\right]$$

$$= \mathbb{E}\left[R_{t+1} \mid S_t = s\right] + \gamma\, \mathbb{E}\left[\underbrace{\mathbb{E}\left[G_{t+1} \mid S_{t+1}\right]}_{\to\, v(S_{t+1})} \mid S_t = s\right]$$

$$\boxed{v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v_{s'}} \longleftarrow \text{Bellman equati}$$

In matrix form:

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

$$\underline{v} = \underline{\mathcal{R}} + \gamma\, \underline{\underline{P}}\, \underline{v} \longrightarrow \text{linear matrix equation}$$

$$\underline{v} = \left(\underline{\underline{I}} - \gamma \underline{\underline{P}}\right)^{-1} \underline{\mathcal{R}} \longrightarrow \text{matrix inversion in } O(N^3)$$

where $N = |\mathcal{S}|$.

Impractical for large N : alternative methods

↳ dynamical programming

↳ Monte Carlo evaluation

Markov decision process

The agent collecting reward can perform actions that will biased future reward collection. This means the transition probabilities of the environment can be affected by the agent's decision.

Formally, a Markov decision process is given by $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$

* $\mathcal{S}$ is a finite set of states.
* $\mathcal{A}$ is a finite set of actions.
* P is a state transition matrix: $P^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$
* R is a reward function: $\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
* γ is a discount factor: $0 \leq \gamma \leq 1$.

Action policy

A policy π is a distribution of actions on $\mathcal{A}$ given states in $\mathcal{S}$.

$\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s]$

↳ policy may be stochastic or deterministic (here deterministic is enough)

↳ policy may be stationary or time dependent (here stationary is enough)

Restriction to stationary deterministic policies can be justified by the memoryless setting of finite state reward Markov process.

Remark about Markovianity

Given a MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and fixed policy π, the joint sequence of states and rewards (S_t, R_t) defines a Markov chain. Justification via averaging over policy actions:

$$P^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) P^a_{s,s'} \quad \leftarrow \text{transition of environment process}$$

$$R^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) R^a_{s,s'} \quad \leftarrow \text{transition of reward process}$$

Action-value function

The action-value function of a MDP measures the expected return of taking the action a in state s under policy π :

$$q_\pi(s,a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right] \leftarrow \text{policy dependent!}$$

Related state-value function:

$$v_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right] = \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[G_t \mid S_t = s, A_t\right]\right]$$

$$= \mathbb{E}_\pi\left[q_\pi(s, A_t)\right] = \sum_{a \in \mathcal{A}} \pi(a,s)\, q_\pi(s,a)$$

Reciprocally:

$$q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$= \mathbb{E}_\pi\left[R_{t+1} \mid S_t = s, A_t = a\right] + \gamma\, \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1}\right] \mid S_t = s, A_t = a\right]$$

$$= \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_\pi(s')$$

Bellman equation:

$$\boxed{v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_\pi(s') \right)}$$

Matrix equation:

$$\left.\begin{aligned} \underline{v_\pi} &= \underline{\mathcal{R}}^\pi + \gamma\, \underline{\underline{P}}^\pi\, \underline{v_\pi} \\ \underline{v_\pi} &= \left(\underline{\underline{I}} - \gamma\, \underline{\underline{P}}^\pi\right)^{-1} \underline{\mathcal{R}}^\pi \end{aligned}\right] \text{ linear equation}$$

Optimal functions

The optimal value function $v_*(s)$ is obtained by maximization over all policies π :

$$v_*(s) = \max_\pi v_\pi(s) \qquad \pi \in \text{convex, simplex space of distributions}$$

The optimal action-value function is similarly defined as

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

A MDP is solved when the optimal value v_* is known. The optimal value function specifies the best possible performance of the MDP.

Notion of optimality

Partial ordering on policies $\pi \geqslant \pi' \Leftrightarrow v_\pi(s) \geqslant v_{\pi'}(s)$ for all s.

Theorem:

For all MDP, there is an optimal policy π_* such that $\pi_* \geqslant \pi$ for any π. All optimal policies achieve the optimal value function $v_*(s) = v_{\pi_*}(s)$. All optimal policies achieve the optimal action-value function $q_*(s,a) = q_{\pi_*}(s,a)$.

Idea of the proof

Bellman's optimality principle:

"An optimal policy has the property that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision".

More formally, the optimal value function is recursively determined by the Bellman optimality equations. Intuitively, we have

1) $v_*(s) = \max_a q_*(s,a)$ ← value given most beneficial choice

2) $q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s')$ ← value of action a given optimal value

$$1) + 2) \Rightarrow \boxed{v_*(s) = \max_a \left(\mathcal{R}_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \right)}$$

Bellman optimality equation

Similarly: $q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} q_*(s',a')$.

Proof of Bellman optimal equation

For any policy $\pi = (\pi_0, \pi')$ ← not necessarily stationary

↳ can be assumed deterministic $\pi_0 = \delta_a$.

$$v_*(s) = \max_\pi \mathbb{E}_\pi [G_t | S_t = s] = \max_\pi \mathbb{E}_\pi \left[\sum_{k \geqslant 0} \gamma^k R_{t+k+1} \middle| S_t = s \right]$$

$$v_*(s) = \max_{\pi} \mathbb{E}_\pi \left[R_{t+1} + \gamma \Big(\sum_{k \geq 0} \gamma^k R_{t+k+2} \Big) \,\Big|\, S_t = s \right]$$

$$= \max_{\pi} \mathbb{E}_\pi \left[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s \right]$$

$$= \max_{(\pi_0, \pi')} \left\{ \mathbb{E}_{\pi_0} \left[R_{t+1} \,|\, S_t = s \right] + \gamma \mathbb{E}_{\pi_0} \left[\mathbb{E}_{\pi'} \left[G_{t+1} \,|\, S_{t+1} \right] \,\big|\, S_t = s \right] \right\}$$

$$= \max_{\pi_0} \left\{ \mathcal{R}_s^{\pi_0} + \gamma \mathbb{E}_{\pi_0} \Big[\underbrace{\max_{\pi'} \mathbb{E}_{\pi'} \left[G_{t+1} \,|\, S_{t+1} \right]}_{v_*(S_{t+1})} \,\Big|\, S_t = s \Big] \right\}$$

$$= \max_{a} \left\{ \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_*(s') \right\}$$

<u>Proof of theorem:</u>

Interpreting Bellman optimality equation as a fixed point equation suggests considering the Bellman operator:

$$T : \mathbb{R}^N \longrightarrow \mathbb{R}^N, \quad N = |\mathcal{S}|$$

$$TV(s) = \max_{\pi} \left\{ \pi(a|s) \Big(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V(s') \Big) \right\}$$

| π can be assumed to be deterministic $\pi = \delta_a$

The operator T is monotone, i.e, if $V \leq V'$ then $TV \leq TV'$.

More importantly T is a contraction for the L_∞ norm on $\mathbb{R}^N$.

Indeed:

$$|TV(s) - TV'(s)| = \Big| \max_{a \in A} \Big[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V(s') \Big] - \max_{a' \in A} \Big[\mathcal{R}_s^{a'} + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^{a'} V'(s') \Big]$$

$$(*) \quad \leq \max_{a \in A} \Big| \cancel{\mathcal{R}_s^a} + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V(s') - \cancel{\mathcal{R}_s^a} - \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V'(s') \Big|$$

$$\leq \underset{\substack{\uparrow \\ <1}}{\gamma} \|V - V'\|_\infty \underbrace{\max_{a \in A} \Big| \sum_{s' \in \mathcal{S}} P_{ss'}^a \Big|}_{\leq 1}$$

(*) from $\left| \max_a f(a) - \max_{a'} g(a') \right| \leq \max_a |f(a) - g(a)|$

Justification:

$$\max_{a} f(a) - \max_{a'} g(a) = f(a^*) - \max_{a'} g(a')$$
$$\leq f(a^*) - g(a^*)$$
$$\leq \max_{a} |f(a) - g(a)| .$$

The contraction property of T allows one to use the Banach fixed point theorem on $(\mathbb{R}^N, \| \|_\infty)$ to show that there is a unique solution v^*, the optimal value function,

Moreover: $\lim_{k \to +\infty} T^k v = v^*$.