# STATE AGGREGATION for MDPs

Olivier Tsemogne Alexandre Reiffers-Masson

RAMONaaS

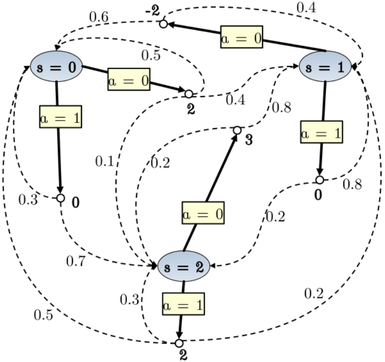
June 26, 2024

Markov decision process

Dimension Reduction

State aggregation

# An example of Markov Decision Process




3 states: $\mathcal{S} = \{0, 1, 2\}$

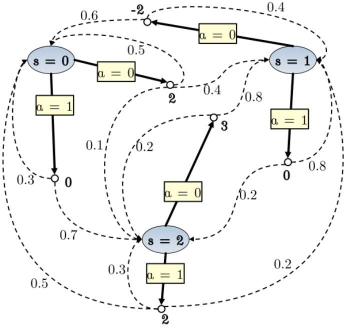
2 actions: $\mathcal{A} = \{0, 1\}$

Reward process: $\mathbf{R} = \begin{bmatrix} 2 & 0 \\ -2 & 0 \\ 3 & 2 \end{bmatrix}$

Transition process:

$$\mathbf{P} = \left[ \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0 & 0.7 \end{bmatrix} \quad \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0 & 0.8 & 0.2 \end{bmatrix} \quad \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.5 & 0.2 & 0.3 \end{bmatrix} \right]$$

# Solution: Policy, Value Function




Policy = action to perform at each state: $s \overset{\pi}{\longmapsto} a = \pi(s)$

Value = long term payoff:

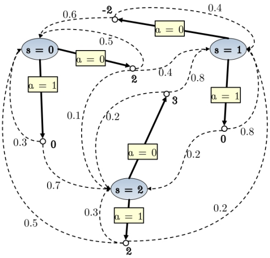
$$s \overset{V^\pi}{\longmapsto} \mathbb{E}\left[\sum_t \gamma^t R^{(t)} \middle| S_0 = s\right]$$

Important property of the value function: It is the fixed point of operator $V \overset{\mathcal{B}^\pi}{\longmapsto} \mathcal{B}^\pi V$ defined by $[\mathcal{B}^\pi V](s) = R_s^{\pi(s)} + \gamma \left\langle V, P_s^{\pi(s)} \right\rangle$

So it is the limit of sequence $V_{n+1} = \mathcal{B}^\pi V_n$

Value of $\pi = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ with $\gamma = 0.9$: $V^\pi = \begin{bmatrix} 7.49 \\ -8.20 \\ -7.20 \end{bmatrix}$.

# Optimal Solution




Objective: maximize $V^{\pi}(s)$ at each $s$. Opimal policy: $\pi^*$; Optimal value function: $V^* = V^{\pi^*}$

$V^*$ is the solution of the **Bellman equation** $\mathcal{B}V = V$ where $\mathcal{B}$ is the optimal Bellman operator defined by:

$$[\mathcal{B}V](s) = \max_{a \in \mathcal{A}} \left( R_s^a + \gamma \langle V, P_s^a \rangle \right).$$

Resolution:

$V_{n+1} \leftarrow \mathcal{B}V_n$

until $\|V_{n+1} - V_n\|_\infty$ is small enough

In our example:

$$\pi^* = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ for } \gamma = 0.4 \qquad \pi^* = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \text{ for } \gamma = 0.9$$

# Scalability of the Value Iteration Algorithm

Most of the time, solving an MDP consists in solving the following **Bellman equation** in $V$:

$$V(s) = \max_{a \in \mathcal{A}} \left( R_s^a + \gamma \langle V, P_s^a \rangle \right), \; \forall s \in \mathcal{S}.$$

**Issue**: The difficulty of solving such equation is exponential in $|\mathcal{S}|$.

**One strategy:** Use solutions of small MDPs to solve large MDPs. More precisely, we can think of two possibilities:

1. Use similarity between the initial problem with smaller problems
2. Transform the large problem in a smaller one

In our case

Markov decision process

Dimension Reduction

State aggregation

# Basic tools: Optimal Transport [Kantorovich, 1942]

- We have 1 compact metric set $\mathcal{S}$ and 2 probability distributions $P$ and $Q$.
- **Problem:** How to optimally modify the mass distribution from $P$ to $Q$ ?
- **Answer:** Consider
  1. the *cost* $c\left(x, y\right)$ for transporting a unit of mass from $x$ to $y$.
  2. the translocation of masses as a probability $\delta$ on $\mathcal{S} \times \mathcal{S}$ for which $P$ and $Q$ are marginals.
  3. the *translocation cost* to be

$$K_c\left(P, Q\right) = \inf_{\substack{\delta \in \Delta(\mathcal{S}\times\mathcal{S}) \\ P \text{ and } Q \text{ are marginals of } \delta}} \int_{\mathcal{S}\times\mathcal{S}} c\left(x, y\right) \mathrm{d}\delta\left(x, y\right).$$

When the cost $c(x, y)$ is a distance, $K_c(P, Q)$ is called the
**Wasserstein-1 distance or Kantorovich-Rubinstein metric.**

# Distance between two states

Consider $\mathcal{M}_1 = (\mathcal{S}_1, \mathcal{A}, \mathbf{P}_1, \mathbf{R}_1, \gamma)$ and $\mathcal{M}_2 = (\mathcal{S}_2, \mathcal{A}, \mathbf{P}_2, \mathbf{R}_2, \gamma)$

- Set an ***initial*** distance $d^{init}(s_1, s_2)$ ***between states*** $s_1 \in \mathcal{S}_1$ and $s_2 \in \mathcal{S}_2$, for all $s_1, s_2$.
- We define the distance between $s_1$, $s_2$ under each action $a$ to be equal to:

$$d^a(s_1, s_2) = c_{\mathrm{R}} \left| (R_1)^a_{s_1} - (R_2)^a_{s_2} \right| + c_{\mathrm{T}} K_{d^{init}} \left( (P_1)^a_{s_1}, (P_2)^a_{s_2} \right).$$

- The ***final distance between states*** is the worst case similarity:

$$d^{final}(s_1, s_2) = \max_{a \in \mathcal{A}} \left( d^a(s_1, s_2) \right), \ \forall s_1, \ s_2.$$

# Distance between two MDPs

Define the *distance* ***between state spaces*** as the optimal translocation from $\mathcal{S}_1$ to $\mathcal{S}_2$ where the cost per unit of mass is $d^{final}(s_1, s_2)$ and the mass is distributed uniformly in each state space:

$$
\begin{aligned}
\Psi(\mathcal{M}_1, \mathcal{M}_2) = \quad & \min_{u \in \mathbb{R}^{\mathcal{S}_1 \times \mathcal{S}_2}} \quad \sum_{(s_1, s_2) \in \mathcal{S}_1 \times \mathcal{S}_2} u_{s_1, s_2} d^{final}(s_1, s_2) \\
& \text{subject to} \quad \sum_{s_2 \in \mathcal{S}_2} u_{s_1, s_2} = \frac{1}{|\mathcal{S}_1|}, \\
& \qquad\qquad\quad \sum_{s_1 \in \mathcal{S}_1} u_{s_1, s_2} = \frac{1}{|\mathcal{S}_2|}, \\
& \qquad\qquad\quad u_{s_1, s_2} \geqslant 0.
\end{aligned}
$$

# Solution Transfer [Song et al., 2016]

**Input:** Large scale MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, \gamma)$, Small MDPs $\mathcal{M}_n = (\mathcal{S}_n, \mathcal{A}, \mathbf{P}_n, \mathbf{R}_n, \gamma)$, for all $n = 1, \ldots, N$ with the optimal $Q$-values of the $N$ small MDPs

1. Consider some metrics $d_n^{init}(s, s_n)$ between states of $\mathcal{M}$ and the small MDPs.
2. Infer the distance between $\mathcal{M}, \mathcal{M}_n$ ($\Psi(\mathcal{M}, \mathcal{M}_n)$) and coefficients for MDP similarity [1] (associated $(u_{s,s_n}^{(n)})_{s,s_n}$), for all $n$.
3. Aggregate the optimal $Q$-values:

$$Q(s,a) = \frac{1}{N} \sum_{n=1}^{N} \sum_{s_n \in \mathcal{S}_n} \frac{u_{s,s_n}^{(n)}}{\sum_{s'} u_{s',s_n}^{(n)}} Q_n^*(s_n, a).$$

[1] See [García et al., 2022] for other similarity metrics.

Markov decision process

Dimension Reduction

State aggregation

# State Space Abstraction with Uniform Weight Distribution [Ferns et al., 2004]

**Input:** Large scale MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, \gamma)$ and a finite set $\mathcal{U}$ that partitions $\mathcal{S}$. Moreover, we have an aggregation function $\phi\colon \mathcal{S} \longrightarrow \mathcal{U}$.

We have the following three steps:

1. **Infer** the aggregated reward and aggregated transition matrix on $\mathcal{U}$ from $\mathcal{M}$ on $\mathcal{S}$:

$$\overline{R}^a_u = \tfrac{1}{|\phi^{-1}(u)|} \sum_{s \in \phi^{-1}(u)} R^a_s,\ \forall u, a,$$

$$\overline{P}^a_{u,u'} = \tfrac{1}{|\phi^{-1}(u)|} \sum_{s \in \phi^{-1}(u)} \sum_{s' \in \phi^{-1}(u')} P^a_{s,s'},\ \forall u, u', a.$$

   The aggregated MDP is given by $\overline{\mathcal{M}} = \left(\mathcal{U}, \mathcal{A}, \overline{\mathbf{P}}, \overline{\mathbf{R}}, \gamma\right)$.

2. **Solve** the aggregated MDP and get the optimal solution $\mu^*$
3. **Return** the extrapolation of the optimal aggregated control:

$$\pi^*(s) = \mu^*(\phi(s)).$$

# Examples

- **Model-irrelevance aggregation:**

$$\phi\left(s_1\right) = \phi\left(s_2\right) \Longleftrightarrow \left\{ \begin{array}{ccc} R^a_{s_1} & = & R^a_{s_2} \\ P^a_{s_1,\cdot} & = & P^a_{s_2,\cdot} \end{array} \right. \forall a$$

- $Q^\pi$**-irrelevance aggregation:**

$$\phi\left(s_1\right) = \phi\left(s_2\right) \Longleftrightarrow Q^\pi\left(s_1, a\right) = Q^\pi\left(s_2, a\right) \forall u, a$$

Some other noticeable abstraction techniques in [Li et al., 2006].

# An Upper Bound of the Error

We need to solve a large scale MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, \gamma)$. So far:

- We have built an abstract version
- We know the abstract solution $\mu^*$ and its extrapolation $\widetilde{\pi}^* = \mu^* \circ \phi$

**Question: What is the difference between the optimal value $V^*$ and the value $V^{\widetilde{\pi}^*}$ of $\widetilde{\pi}^*$?**

**Answer** [Bozkurt et al., 2023]:

$$\left\| V^* - V^{\widetilde{\pi}^*} \right\|_\infty \leqslant \frac{2}{1-\gamma} \Delta^{\max} v^*,$$

where $v^* =$ approximated value function, and

$$\Delta^{\mathsf{max}} V = \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \left[\mathcal{B}^a V\right](s) - \left[\widetilde{\mathcal{B}}^a V\right](s) \right|.$$

# Upper Bound under Assumption of Lipschitz Continuity

Theorem

*Upper bound performance:*

1. ***Dirac measure***

$$\left\| V^* - V^{\widetilde{\pi}^*} \right\|_\infty \leqslant \frac{\sup\limits_{s \in \mathcal{S}} d_{\mathcal{S}}\left(s, \widehat{s}\right)}{1-\gamma} \left( L_{\mathbf{R}} + \gamma L_{\mathbf{P}} \left\| V^* \right\|_{\mathrm{L}} + \frac{1+\gamma}{1-\gamma} \left\| V^* \right\|_{\mathrm{L}} \right),$$

   *where* $L_{\mathbf{R}}$, $L_{\mathbf{P}}$, $\left\| V^* \right\|_{\mathrm{L}} =$ *Lipschitz coefficients,* $\sup\limits_{s \in \mathcal{S}} d_{\mathcal{S}}\left(s, \widehat{s}\right) =$ *maximum diameter.*

2. ***General case:***

$$\left\| V^* - V^{\widetilde{\pi}^*} \right\|_\infty \leqslant \frac{2\delta_{\phi,\omega}}{\left(1-\gamma\right)^2} \left( L_{\mathbf{R}} + \gamma L_{\mathbf{P}} \left\| V^* \right\|_{\mathrm{L}} \right),$$

   *with* $\delta_{\phi,\omega} =$ *maximum mean diameter.*

# References


Bozkurt, B., Mahajan, A., Nayyar, A., and Ouyang, Y. (2023).
Weighted-norm bounds on model approximation in mdps with unbounded per-step cost.
In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7817–7823. IEEE.

Ferns, N., Panangaden, P., and Precup, D. (2004).
Metrics for finite markov decision processes.
In *UAI*, volume 4, pages 162–169.

García, J., Visús, Á., and Fernández, F. (2022).
A taxonomy for similarity metrics between markov decision processes.
*Machine Learning*, 111(11):4217–4247.

Kantorovich, L. V. (1942).
On the translocation of masses.
*Dokl. Akad. Nauk. USSR (N.S.)*, 37:199–201.

Li, L., Walsh, T. J., and Littman, M. L. (2006).
Towards a unified theory of state abstraction for mdps.
*AI&M*, 1(2):3.

Song, J., Gao, Y., Wang, H., and An, B. (2016).
Measuring the distance between finite markov decision processes.
In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 468–476.