What is the state-representation problem?

Shibhansh Dohare
RLAI, University of Alberta

Problem setting

1. World/Environment is a stationary MDP

Problem setting

1. World/Environment is a stationary MDP
2. Continuing setting

Problem setting

1. World/Environment is a stationary MDP
2. Continuing setting
3. The environment state is only partially observable; the agent receives O_t at every time step

Problem setting

1. World/Environment is a stationary MDP
2. Continuing setting
3. The environment state is only partially observable; the agent receives O_t at every time step
4. **History** refers to an initial portion of the trajectory up to an observation at time step *t*, i.e. $H_t = A_0, O_1, \ldots, A_{t-1}, O_t$

Problem setting

1. World/Environment is a stationary MDP
2. Continuing setting
3. The environment state is only partially observable; the agent receives O_t at every time step
4. **History** refers to an initial portion of the trajectory up to an observation at time step *t*, i.e. $H_t = A_0, O_1, \ldots, A_{t-1}, O_t$
5. A **Future** refers to a possible sequence of observations and actions, i.e. - $T_{t+1} = O_{t+1}, A_{t+1}, O_{t+2}, A_{t+2}, \ldots$

Problem setting

1. World/Environment is a stationary MDP
2. Continuing setting
3. The environment state is only partially observable; the agent receives O_t at every time step
4. **History** refers to an initial portion of the trajectory up to an observation at time step *t*, i.e. $H_t = A_0, O_1, \ldots, A_{t-1}, O_t$
5. A **Future** refers to a possible sequence of observations and actions, i.e. - $T_{t+1} = O_{t+1}, A_{t+1}, O_{t+2}, A_{t+2}, \ldots$

Note - I'm not considering non-stationary MDPs, as any non-stationary MDP can be modelled as a partially observable MDP and vice-versa.

The various kinds of states

1. Environment state - The state of the MDP
2. Information state -
3. Markov state -
4. Agent state -

Information state

It is an equivalence class over histories for which all futures have the same probability of happening, i.e. two histories *h* and *h'* belong to the same information state if and only if

$$Pr(F = \tau | H_t = h, A_t = a) = Pr(F = \tau | H_t = h', A_t = a) \forall \tau, a$$

Information state

It is an equivalence class over histories for which all futures have the same probability of happening, i.e. two histories *h* and *h'* belong to the same information state if and only if

$$Pr(F = \tau | H_t = h, A_t = a) = Pr(F = \tau | H_t = h', A_t = a) \forall \tau, a$$

The information states form an MDP[1] -

- At every time step, the agent is in one of the information states
- The transition from one information state to the next is conditionally independent of past information states given the current information state

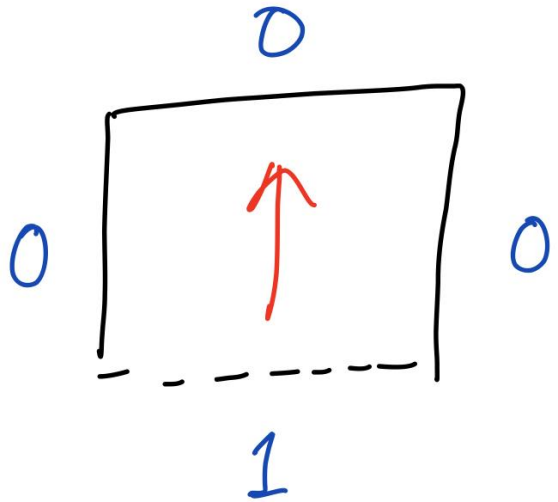
[1] Cosma Rohilla Shalizi, James P. Crutchfield, *Computational Mechanics: Pattern and Prediction, Structure and Simplicity,* Journal of Statistical Physics, 1572-9613, 2001

An example

A four-state MDP, red arrow is the agent -

Actions - *Left (L)*, *Right (R), Spin(S)*

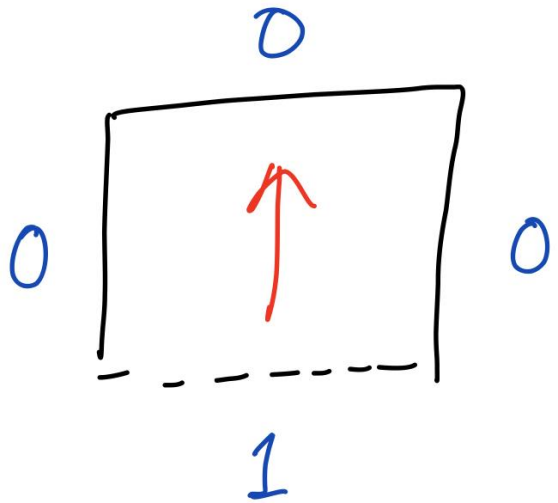
- Agent receives observations ‘0’ or ‘1’
- *Right* rotates the agent to the right (of the agent) by 90 degrees
- *Left* rotates the agent to the left by 90 degrees
- *Spin* results with the agent in a random state



An example

Information states -

- ...1 | ...S0R1 | ...R0R1 |
- ...1R0 | ...S0L0L0 |
- ...1L0 | ...S0R0R0 |
- ...1R0R0 | ...1L0L0 |
- ...S0 |
- ...S0R0 | ...S0L0R0
- ...S0L0 | ...S0R0L0

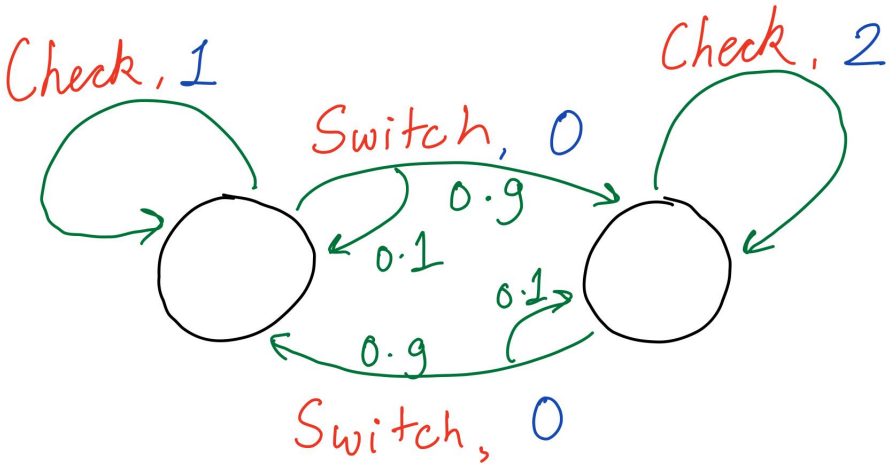


Another example

A two-state MDP -

Actions - *Check (C)*, *Switch (S)*

- *Check* gives observations ‘1’ or ‘2’
- *Switch* transitions to the other state with probability 0.9

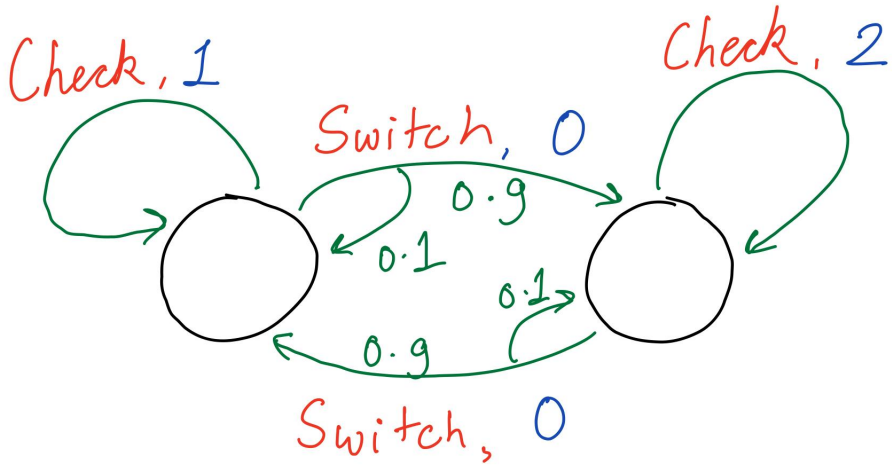


Another example

Information states -

- … C1 |
- … C2 |
- … C2S0 |
- … C2S0S0 |
- … C2S0S0S0 |

There are infinite information states



Additional Constraints - The aperture principle

The environment is much more complex than the agent - i.e. the computation/memory required to update the environment state is much larger than the computation/memory available to the agent.

The observations that the agent receives at every time step are much smaller than the memory available to the agent.

Information state

Directly approximating information state might not be a good idea

Information state

Directly approximating information state might not be a good idea

- They capture all the differences about the past that makes a difference in the future for all possible histories.

Information state

Directly approximating information state might not be a good idea

- They capture all the differences about the past that makes a difference in the future for all possible histories.
- It is not possible to determine the information state from a finite amount of history.

Information state

Directly approximating information state might not be a good idea

- They capture all the differences about the past that makes a difference in the future for all possible histories.
- It is not possible to determine the information state from a finite amount of history.
- The algorithms that find probability distribution over information states are computationally expensive [2].

Daniel R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997

Information state

Directly approximating information state might not be a good idea

- They capture all the differences about the past that makes a difference in the future for all possible histories.
- It is not possible to determine the information state from a finite amount of history.
- The algorithms that find probability distribution over information states are computationally expensive [2].
- To the best of my knowledge, there is no work on MDPs with infinite hidden states.

Daniel R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997

The various kinds of states

1. Environment state - The state of the MDP
2. Information state - It is an equivalence class over histories for which all futures have the same probability of happening
3. Markov state -
4. Agent state -

Markov state

Markov property - It is a property of a function *f*; the function has this property when any two histories *h* and *h'* that are mapped to the same state also have the same probability distribution over futures -

$$f(h) = f(h') \Rightarrow Pr(T = \tau | H_t = h, A_t = a) = Pr(T = \tau | H_t = h', A_t = a) \forall \tau, a$$

Markov state

Markov property - It is a property of a function *f*; the function has this property when any two histories *h* and *h'* that are mapped to the same state also have the same probability distribution over futures -

$$f(h) = f(h') \Rightarrow Pr(T = \tau | H_t = h, A_t = a) = Pr(T = \tau | H_t = h', A_t = a) \forall \tau, a$$

Markov state is the output of a Markov function

Markov state

Markov property - It is a property of a function *f*; the function has this property when any two histories *h* and *h'* that are mapped to the same state also have the same probability distribution over futures -

$$f(h) = f(h') \Rightarrow Pr(T = \tau | H_t = h, A_t = a) = Pr(T = \tau | H_t = h', A_t = a) \forall \tau, a$$

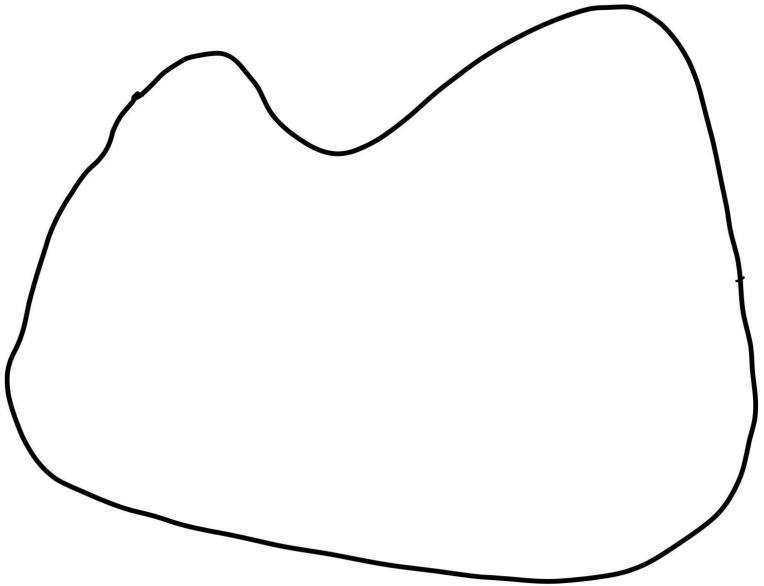
Markov state is the output of a Markov function

Note - For a function *g*, that maps histories to information states,

$$g(h) = g(h') \Leftrightarrow Pr(T = \tau | H_t = h, A_t = a) = Pr(T = \tau | H_t = h', A_t = a) \forall \tau, a$$

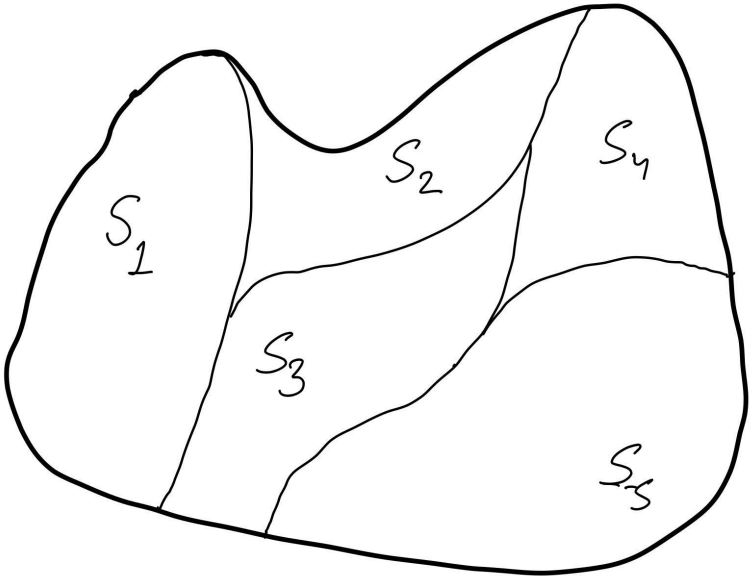
Markov State

- Set of all histories



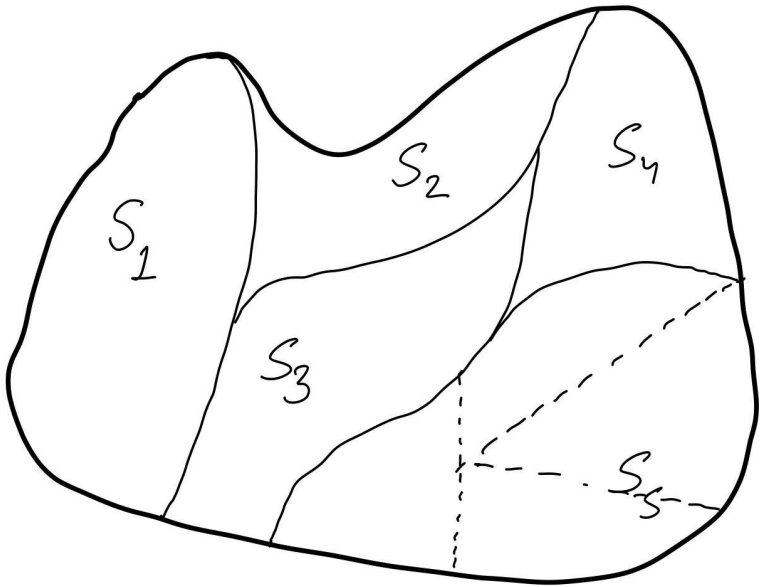
Markov State

- Set of all histories
- Partitions created by solid lines represent information states



Markov State

- Set of all histories
- Partitions created by solid lines represent information states
- Markov functions further divide the information states into Markov states, represented by dashed lines



A trivial example of Markov state

A function that maps all histories to a different state is a Markov function.

But this is a very bad (maybe worst?) Markov function.

A trivial example of Markov state

A function that maps all histories to a different state is a Markov function.

But this is a very bad (maybe worst?) Markov function.

- Information states hide all differences in histories that don’t affect the future
- We want our Markov functions to hide at-least some differences in histories that don’t affect the future
- But this trivial Markov function doesn’t hide any useless information.

The various kinds of states

1. Environment state - The state of the MDP
2. Information state - It is an equivalence class over histories for which all futures have the same probability of happening
3. Markov state - Probability distribution over all futures is conditionally independent of all past observations given the state
4. Agent state -

The various kinds of states

1. Environment state - The state of the MDP
2. Information state - It is an equivalence class over histories for which all futures have the same probability of happening
3. Markov state - Probability distribution over all futures is conditionally independent of all past observations given the state
4. Agent state - An approximate Markov state that the agent uses to control and predict the future

Some thoughts about solution

The solution methods from the fully observable case might not transfer to this setting - a feed-forward network vs a recurrent solution.

Some thoughts about solution

The solution methods from the fully observable case might not transfer to this setting - a feed-forward network vs a recurrent solution.

If the state-update function doesn’t hide any information, i.e. it doesn’t map histories with the same distribution over futures to the same state, then it is not helping.

Some thoughts about solution

The solution methods from the fully observable case might not transfer to this setting - a feed-forward network vs a recurrent solution.

If the state-update function doesn’t hide any information, i.e. it doesn’t map histories with the same distribution over futures to the same state, then it is not helping.

If the state-update function hides useful information, i.e. it maps histories with different distributions over futures to the same state, then it may be catastrophic, or it may help.

Relation to Policy gradient

If the agent has a perfect Markov state - then the best policy will be deterministic

But it may never happen, so it is better to use policy gradient based methods, that can handle non-Markov states better than value-based methods[3]

[3] Richard Sutton and Andrew Barto. Reinforcement Learning: An Introduction. MIT Press,2017.

Relation to tracking/meta-learning

What is tracking/meta-learning?

- The best value function/policy changes with time because the world appears non-stationary
- SGD without decaying learning rate, IDBD

Again, imperfections in the state construction are what makes tracking useful; it doesn’t make sense to track the value function/policy if the state is Markov.

Questions?

State of the agent

Set of all the variables that change inside the agent that and determine the future of the agent.