Exact Reduction of Huge Action Spaces in General Reinforcement Learning

Sultan J. Majeed[1]   Marcus Hutter[2]

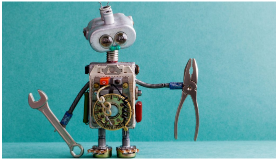
[1,2]Australian National University

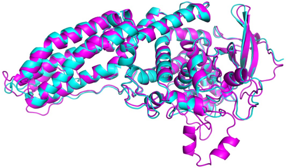
[2]Google DeepMind

35th AAAI Conference of Artificial Intelligence, February 2021

Introduction

- Many reinforcement learning (RL) problems have huge action-spaces.





Examples: Robotics, Protein Folding, and StarCraft.[1]

[1] Image credit: Createdigital, MIT Technology Review, Full-stack Feed

Introduction (Cont.)

- Observations $\neq$ States, i.e. most problems are non-Markovian.
- Need to keep (parts of) the history to define the “state”.

Examples: Self-driving Cars, Montezuma’s Revenge, and Minecraft.[2]

[2] Image credit: Yahoo Finance, Medium, Minecraft Wiki

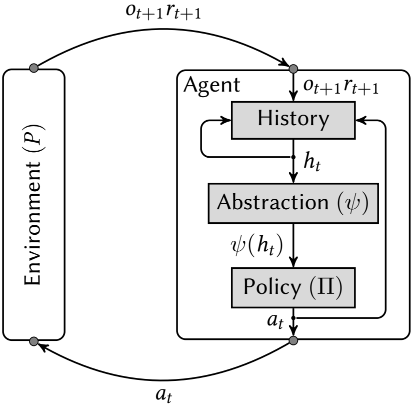
Research Question

Is it possible, in theory, to reduce any (history-based) problem with a huge action-space to a reasonably sized state-action space MDP model?

This work: Yes!

History-based RL with Abstraction

- At time t, the agent takes an action $a_t \in \mathscr{A}$.
- The environment dispatches an observation-reward pair $o_{t+1}r_{t+1} \in \mathscr{O} \times \mathscr{R}$.
- The updated history is $h_t := h_{t-1}a_{t-1}o_tr_t \in \mathscr{H}$.
- The abstraction $\psi : \mathscr{H} \to \mathscr{S}$ provides a (sufficient) statistics of the history.
- The agent selects actions through a policy $\Pi : \mathscr{S} \to \triangle(\mathscr{A})$.



The agent-environment interaction.

Action-value Uniform Abstractions

Definition (ε-Q-uniform abstraction)

An abstraction function $\psi : \mathscr{H} \to \mathscr{S}$ is an ε-Q-uniform abstraction if for any $h, \dot{h} \in \mathscr{H}$ and all $a \in \mathscr{A}$ we have

$$\left(\psi(h) = \psi(\dot{h})\right) \implies \left|Q^*(h, a) - Q^*(\dot{h}, a)\right| \leq \varepsilon$$

where $\mathscr{S}$ is the set of states of the abstraction.

- Q^* is the optimal action-value function.
- An approximation of Q^* can be used in the above definition with an extra error term.

Extreme State Aggregation (ESA)

Theorem (ESA[3])

For every environment P there exists an abstraction and a surrogate-MDP whose optimal policy is an ε-optimal policy for the environment. The size of the surrogate-MDP is bounded (uniformly for any P) by

$$|\mathscr{S}| \leq \left(\frac{2}{\varepsilon(1-\gamma)^3}\right)^{|\mathscr{A}|}$$

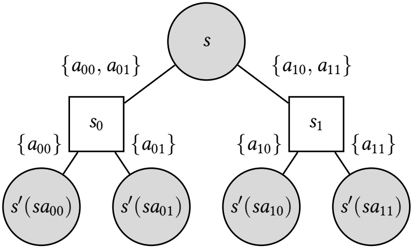
where γ is the discount-factor.

- The size of the abstraction scales exponentially in $|\mathscr{A}|$.
- Not very useful even for medium-sized action-space problems.

[3] Marcus Hutter. "Extreme state aggregation beyond Markov decision processes". In: *Theoretical Computer Science* (2016), pp. 73–91, Theorem 11.

Action Sequentialization in Markovian Environments

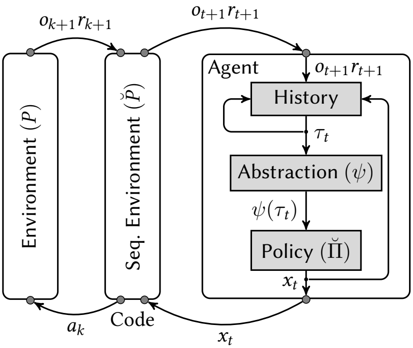
- Let $\mathscr{A} = \{a_{00}, a_{01}, a_{10}, a_{11}\}$.
- $s'(sa)$ denotes the next state s' reached from state s when the agent takes action a.
- The filled circles denote the states of the original MDP.
- The squares denote the added states.



A simple sequentialization example in an MDP.

Sequentialized Environment

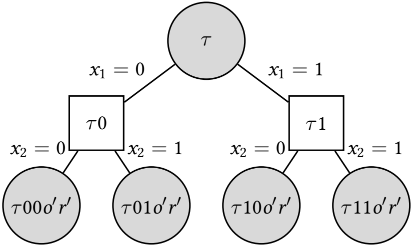
- Biject each $a \in \mathscr{A}$ to a unique $\mathscr{B}$-ary vector $x_1, x_2, \ldots, x_d = \boldsymbol{x} = \mathrm{Code}(a)$ of length d.
- The agent takes a $\mathscr{B}$-ary decision $x_t \in \mathscr{B}$.
- The sequentialized environment ($\mathscr{B}$-ary mock) provides a buffered observation o_{t+1} and reward r_{t+1}.
- Once $\boldsymbol{x} \in \mathscr{B}^d$ decisions are taken, $\mathscr{B}$-ary mock acts on the true environment with $a = \mathrm{Code}^{-1}(\boldsymbol{x}) \in \mathscr{A}$.



The agent-environment interaction through the sequentialization scheme.

Action Sequentialization in History-based Environments

- Let τ be a "sequentialized" history, and $\mathscr{B} = \{0, 1\}$.
- For brevity, the intermediate buffered observation-reward pairs are omitted.
- If P is an MDP then o is a sufficient statistics of τ.



A simple sequentialization/binarization example in a deterministic history-based process.

Sequentialization is Useful

Theorem (Sequentialization preserves Markov property)

If P is an MDP over $\mathscr{O}$, and the observations from the $\mathscr{B}$-ary mock are $\widetilde{\mathscr{O}} := \mathscr{O} \times \cup_{i=0}^{d-1} \mathscr{B}^i$, then sequentialized $\breve{P}$ is also an MDP over $\widetilde{\mathscr{O}}$.

- This construction reduces the action-space at the expense of the state-space from $|\mathscr{O}|$ to $\approx 2|\mathscr{A}| \cdot |\mathscr{O}|$.
- Algorithms which bootstrap can benefit from such sequentialization, e.g. Q-learning.
- Since $\breve{P}$ is also an MDP, the convergence and optimality guarantees in MDPs are carried over to the sequentialized process.

Sequentialization is Useful (Cont.)

Theorem (Sequentialization preserves ε-optimality)

Any $\gamma\varepsilon$-optimal policy of the sequentialized environment is ε-optimal in the original environment.

- It means that we can uplift a near-optimal policy from $\breve{P}$ to P.
- The uplifted policy is guaranteed to be near-optimal.

Binarized ESA

Theorem (Binary ESA)

For every environment there exists an abstraction and a corresponding surrogate-MDP for its binarized version ($\mathscr{B} = \{0, 1\}$) whose optimal policy is ε-optimal for the true environment. The size of the surrogate-MDP is uniformly bounded for every environment as

$$|\mathscr{S}| \lesssim \frac{4\lceil \log_2 |\mathscr{A}| \rceil^6}{\varepsilon^2 (1-\gamma)^6} \quad (\textit{when } \gamma \to 1)$$

- The size of the abstraction scales only logarithmically in $|\mathscr{A}|$.
- The huge action-space problems can be reduced to a binary action-space problem with a significantly improved state-space size.

Key Takeaway

For every RL problem there exists an ε-optimal MDP model with a binary action-space, and the number of states are

$$|\mathscr{S}| \lesssim \frac{4\lceil \log_2 |\mathscr{A}| \rceil^6}{\varepsilon^2 (1-\gamma)^6} \quad (\textit{when } \gamma \to 1)$$

Further Questions

Thanks for your attention!

Reach out to sultan.majeed@anu.edu.au for further questions.