For **every RL problem** there exists a **near-optimal** model with a **binary action-space** and the **number of states** are bounded **uniformly**.

Exact Reduction of Huge Action Spaces in General Reinforcement Learning

Sutan J. Majeed and Marcus Hutter

ANU & Google DeepMind

Introduction

- Many RL problems have **huge action-spaces**.
- **Observations** $\neq$ **States**, i.e. most problems are non-Markovian.
- Need to keep (parts of) the **history** to define the "state".

(So, the key question is ...)

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?*

What About Extreme State Aggregation?

- The ESA framework can provided a **uniform** bound on the **size of the state-space**.
- **But**, the bound scales **exponentially** in the **action-space**.

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

- So, not suitable even for **moderately-sized action-space** problems!
- Is there a way to improve the bound?

(Glad you asked!)

Action Sequentialization is the Key!

- We can sequentialize the decision-making process, e.g. binarization.
- The agent chooses among **two alternatives** at each step.
- The new states are added for these **partial decisions**.

(Wait... You've just blown out the state-space!!)

- *It turns out, the added states are **not necessary** in ESA!*
- *We can use the sequentialized process as a substitute for the true environment.*
- *Then ESA provides the **model existence guarantee** for the sequentialized setup.*

(So what? It might not be useful.)

- *We show that the policy of the sequentialized process is **near-optimal** in the true environment.*

(Aha! Now you are talking!)

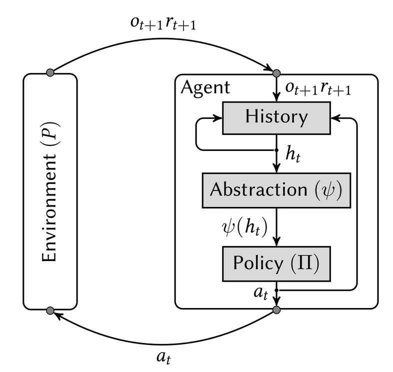
Conclusion

Using the action sequentialization, we were able to prove that, **yes**, there **exists** a map for **every RL problem** with a **reasonably sized** state-action space. The reduced action-space is **binary**, and the **size of the state-space** is

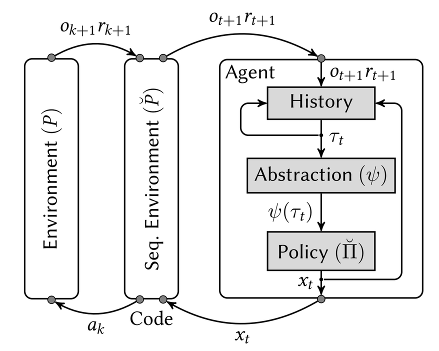
$$|\mathscr{S}| \lesssim \frac{4\lceil \log_2 |\mathscr{A}| \rceil^6}{\varepsilon^2(1-\gamma)^6} \qquad (2)$$

(Can you believe that? A double exponential improvement!)

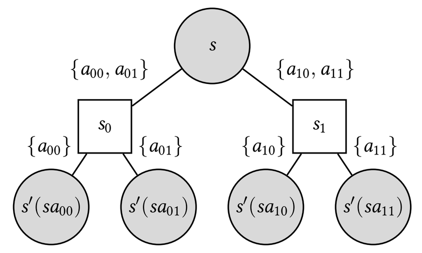
Supplementary Figures



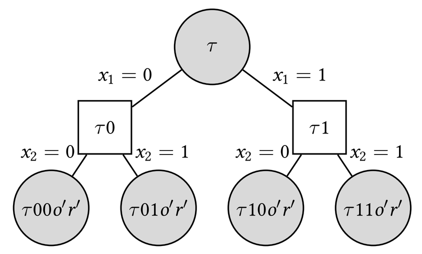
General Reinforcement Learning



GRL with Sequentialized Actions



Action Sequentialization in an MDP



Action Sequentialization in a History-based problem

Full Paper

Australian National University