# Solving *K*-MDPs


Jonathan Ferrer-Mestres
Conservation Decisions
Land and Water, CSIRO

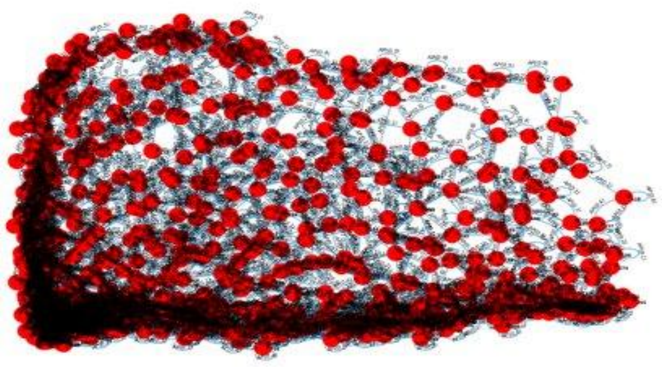
Tom Dietterich
Electrical Engineering and
Computer Science
Oregon State University

Olivier Buffet
INRIA

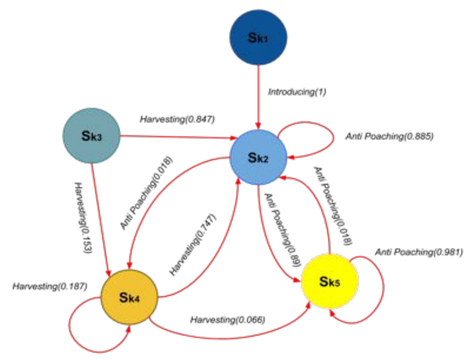
Iadine Chades
Conservation Decisions
Land and Water, CSIRO


Markov decision processes (**MDPs**) are a convenient mathematical model for tackling sequential decision-making problems under uncertainty. MDPs have been applied to domains such as conservation of biodiversity and health. Policies computed for MDPs with thousands of states are in practice difficult to understand by humans. Developing **easy-to-interpret** solutions is crucial to increase uptake of MDP policies. We propose to increase the **interpretability** of MDPs by solving **K-MDPs**, i.e., given an original MDP and a number of states (**K**), generate a reduced state space MDP that **minimizes** the difference between the original and reduced optimal solutions.

Difficult to interpret



We propose to compute MDPs with K states: *K-MDPS*



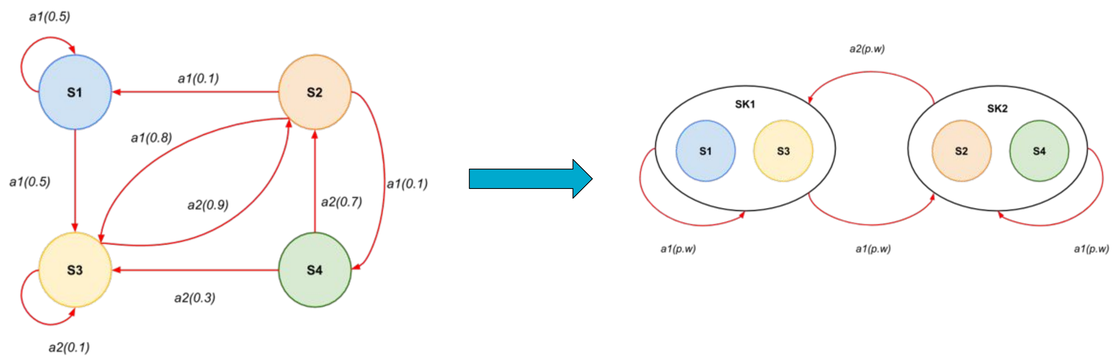

Interpretable solution

## Methods

The aim of this research is to generate a **reduced state space** MDP that **minimizes** the difference between the original optimal MDP value function and the reduced optimal value function.

We solve *K*-MDPs using **state abstractions**. Formally, a *K*-MDP is a tuple $<S_K, A, T_K, r_K, \gamma, \Phi>$, where:

- $S_K$ is the abstract state space;
- $A$ is a set of actions;
- $T_K$ is the abstract *K*-MDP probability transition function;
- $r_K$ is the abstract reward function;
- $\gamma$ is the discount factor;
- $\Phi$ is a mapping function from S to $S_K$.




An optimal solution for a *K*-MDP can be applied to the original MDP using the mapping function $\Phi$:

$$V_{\phi}^{\pi^*_K}(s) = E\left(\sum_{t=0}^{t=H} \gamma^t r(s_t, \pi^*_K(\phi(s_t)))|s_0 = s\right).$$

We formulate the problem of finding the best reduced state space as a **gap minimization** problem:

$$gap^* = \min_{S_K \in P(S), |S_K| \leq K} \max_{s \in S} [V^{\pi^*}(s) - V_{\phi}^{\pi^*_K}(s)]$$

## Proposed algorithms

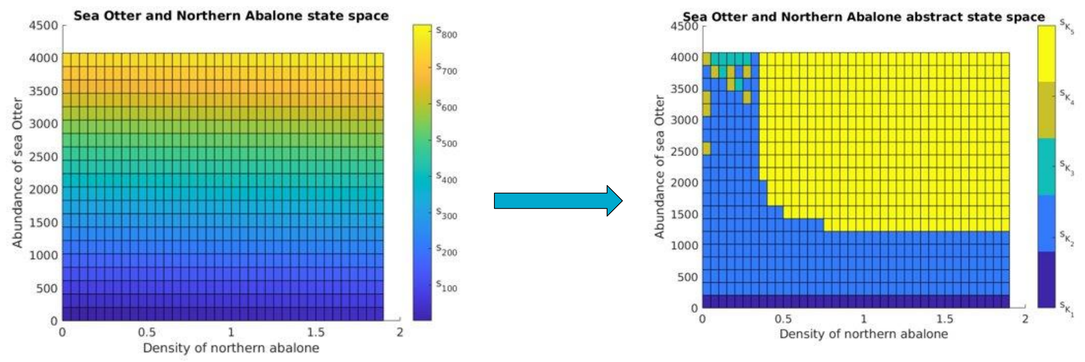
- Our algorithms group states into **bins**.
- If two or more states are in the same bin, they can be **aggregated**.
- **Challenge**: How to define the size of the bin if we want *K* states?

We propose:

- Four algorithms based on **binary search** on $\epsilon$ and $d$.
  - Finding best $\epsilon$ and $d$ that will guarantee the best performance.
- One algorithm based on the **clustering technique** *k*-means++.

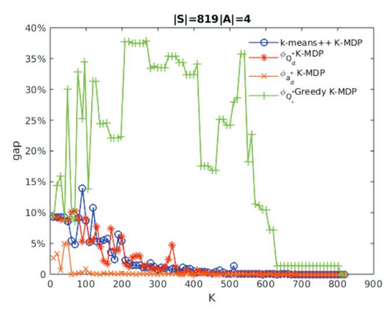
| Algorithm | Function | Predicate | Value Loss | Trans. |
|---|---|---|---|---|
| $\phi_{Q^*_\epsilon} K$-MDP-ILP **new** | $\phi_{Q^*_\epsilon}$ [Abel et al., 2016] | $\max_a \lvert Q^*(i,a) - Q^*(j,a)\rvert \leq \epsilon$ | $\max_{s\in S} V^{\pi^*}(s) - V^{\pi^*_K}_{\phi_{Q^*_\epsilon}}(s) \leq \frac{2\epsilon R_{max}}{(1-\gamma)^2}$ | No |
| $\phi_{Q^*_d} K$-MDP **new** | $\phi_{Q^*_d}$ [Abel et al., 2018] | $\forall a \left\lceil \frac{Q^*(i,a)}{d} \right\rceil = \left\lceil \frac{Q^*(j,a)}{d} \right\rceil$ | $\max_{s\in S} V^{\pi^*}(s) - V^{\pi^*_K}_{\phi_{Q^*_d}}(s) \leq \frac{2dR_{max}}{(1-\gamma)^2}$ | Yes |
| $\phi_{a^*_d} K$-MDP **new** | $\phi_{a^*_d}$ **new** | $a^*_i = a^*_j \wedge \left\lceil \frac{V^*(i)}{d} \right\rceil = \left\lceil \frac{V^*(j)}{d} \right\rceil$ | $\max_{s\in S} V^{\pi^*}(s) - V^{\pi^*_K}_{\phi_{a^*_d}}(s) \leq \frac{2dR_{max}}{(1-\gamma)^2}$ | Yes |
| $\phi_{Q^*_\epsilon}$ Greedy $K$-MDP **new** | $\phi_{Q^*_\epsilon}$ [Abel et al., 2016] | $\max_a \lvert Q^*(i,a) - Q^*(j,a)\rvert \leq \epsilon$ | $\max_{s\in S} V^{\pi^*}(s) - V^{\pi^*_K}_{\phi_{Q^*_\epsilon}}(s) \leq \frac{2\epsilon R_{max}}{(1-\gamma)^2}$ | No |
| $k$-means++ $K$-MDP **new** | - **new** | - | - | Yes |

## Experiments

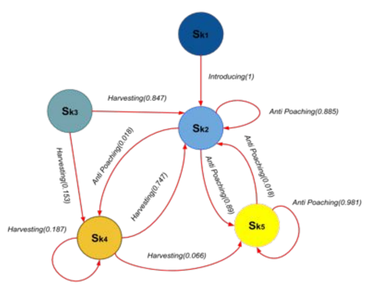



Chades et al., 2012
MDP model with 819 states

Ferrer Mestres et al., 2020
*K*-MDP model with 5 states




Performance of *K*-MDP algorithms.




Reduced policy graph with *K*=5 abstract states.

## Discussion

Our approach aims at increasing **uptake** of MDPs in human operated systems by providing easier to **interpret** models and solutions. We have proposed to solve *K*-MDPs using state abstraction functions and a clustering technique. Future work will test the **interpretability** of proposed *K*-MDP solutions and will address visualization and interpretability challenges in problems with a large number of state variables.

**As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.**

CSIRO. Unlocking a better future for everyone.


FOR FURTHER INFORMATION

Jonathan Ferrer-Mestres
Land and Water, CSIRO
jonathan.ferrermestres@csiro.au
https://www.csiro.au/en/Research/LWF



*REFERENCES*

*[1] Abel, David, David Hershkowitz, and Michael Littman. "Near optimal behavior via approximate state abstraction." ICML. PMLR, 2016.*
*[2] Abel, David, et al. "State abstractions for lifelong reinforcement learning." ICML. 2018.*
*[3] Chadès, Iadine, et al. "General rules for managing and surveying networks of pests, diseases, and endangered species." PNAS 2011.*
*[4] Ferrer-Mestres, Jonathan, et al. "Solving K-MDPs." ICAPS. 2020.*


COLLABORATORS