Solving K-MDPs

Jonathan Ferrer-Mestres[1], **Tom Dietterich**[2], **Olivier Buffet**[3], **Iadine Chades**[1]
[1]**CSIRO**, [2]**Oregon State University**, [3]**INRIA**

Contact: jonathan.ferrermestres@csiro.au

Conservation Decisions Team — Land and Water
www.csiro.au

Markov Decision Process (MDP) and its applications

MDPs have been applied to conservation in bio-diversity:

- to help recover populations under limited resources;
- to control invasive species;
- to manage fisheries;
- to perform adaptive management of natural resources;
- to test behavioral ecology theories.

Example:

- Recovering two endangered species problem [Chades et al., 2012].
- Managing an ecological network of invasive species [Péron et al., 2017].

Fully observable, probabilistic state models. An MDP is specified as $\langle S, A, T, r, \gamma \rangle$, where:

- S is the set of fully observable states;
- A is the set of actions;
- $T : S \times A \times S \to [0, 1]$ is a probabilistic transition function;
- $r : S \times A \to [0, R_{max}]$ is the reward function.
- $\gamma \in [0, 1]$ is a discount factor.

Solutions are functions (policies) mapping states into actions.

Recovering two endangered species: A non-interpretable policy

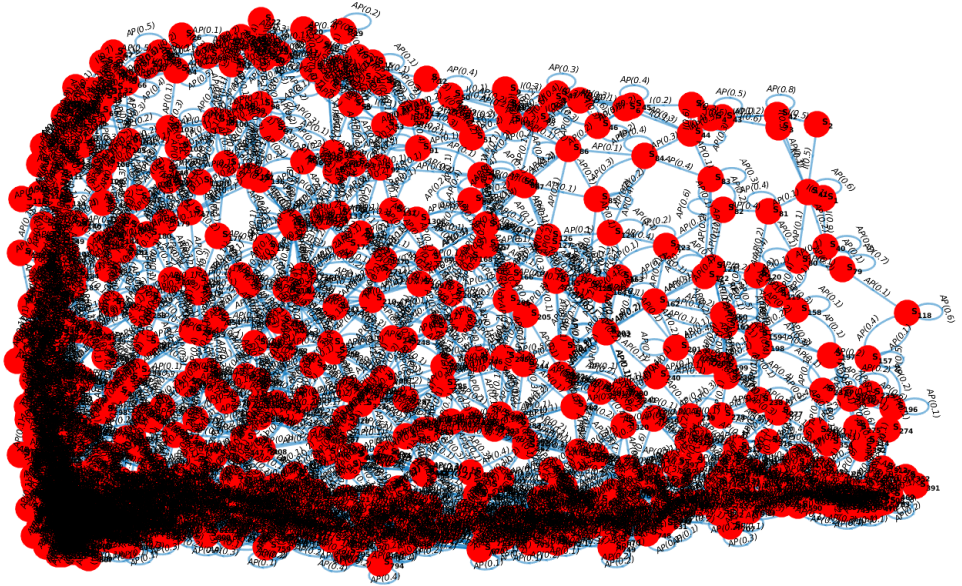


Figure 1: Recovering two endangered species MDP policy graph with 819 states and 4 management actions [Chades et al., 2012].

Building more interpretable models and solutions

Motivation:

- Can we increase models and solutions interpretability in human operated systems?

Our aim:

- Generate a reduced state space MDP that **minimizes** the difference between the original optimal MDP value function and the reduced optimal value function.

We define the problem of solving K-MDPs, given:

- An original MDP.
- A constraint on the number of states (K).

$\times$ We are not trying to solve large MDPs.
$\checkmark$ We are trying to find the most compact MDP model and policy.

K-MDPs: General problem statement

Our contribution:

A K-MDP $M_K = \langle S_K, A, T_K, r_K, \gamma, \phi \rangle$ is an MDP where: the

- S_K is a reduced state set of size at most K;
- A is the original set of actions;
- $T_K : S_K \times A \times S_K \to [0, 1]$ is the probability transition function;
- $r_K : S_K \times A \to [0, R_{max}]$ is the reward function;
- γ is the discount factor;
- ϕ is a mapping function from S to S_K.

An optimal solution for a K-MDP is a policy $\pi^*_K : S_K \to A$ that maximizes the expected sum of discounted rewards.

K-MDPs: General problem statement

Our contribution:

A policy π^*_K can be applied to the original MDP by using the mapping function ϕ, with the associated value function:

$$V_{\phi}^{\pi^*_K}(s) = E\left(\sum_{t=0}^{t=H} \gamma^t r(s_t, \pi^*_K(\phi(s_t)))|s_0 = s\right). \quad (1)$$

We formulate the problem of finding the best reduced state space ($|S_K| \leq K$) as a gap minimization problem:

$$gap^* = \min_{S_K \in P(S), |S_K| \leq K} \max_{s \in S} [V^{\pi^*}(s) - V_{\phi}^{\pi^*_K}(s)], \quad (2)$$

K-MDPs using state abstraction

Lets define a K-MDP as $M_K = \langle S_K, A, T_K, r_K, \gamma, \phi \rangle$:

- $S_K = \{\phi(s) | s \in S\}$ the abstract state space and $\phi : S \to S_K$;
  - The inverse of function $\phi^{-1}(s_K) : S_K \to S$.
- A is the same set of actions as in the original MDP model.
- T_K is the abstract K-MDP probability transition function: $T_K(s_K, a, s'_K) = \sum_{s \in \phi^{-1}(s_K)} \sum_{s' \in \phi^{-1}(s'_K)} T(s, a, s')\omega(s)$, where:
  - $\forall s_K \in S_K, \left(\sum_{s \in \phi^{-1}(s_K)} \omega(s) \right) = 1$ and $\omega(s) \in [0, 1]$
- r_K is the abstract reward function: $r_K(s_K, a) = \sum_{s \in \phi^{-1}(s_K)} r(s, a)\omega(s)$.

Proposed algorithms

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

Table 1: Summary of proposed algorithms.

- Four algorithms based on binary search on ϵ and d:
  - Finding the best ϵ and d that will guarantee the best performance.
- One algorithm based on a clustering technique.

Proposed algorithms: The $\phi_{Q^*_\epsilon} K$-MDP-ILP algorithm

$$\phi_{Q^*_\epsilon}(i) = \phi_{Q^*_\epsilon}(j) \implies \max_a |Q^*(i,a) - Q^*(j,a)| \leq \epsilon. \quad (3)$$

Issue: Non-transitive.

Solution: We solve the problem as a minimum clique node cover problem by building a graph C of possible aggregations [Brigham and Dutton, 1983]:

- $\delta(i,j) = \max_a |Q^*(i,a) - Q^*(j,a)|$, $i, j \in S$ and $a \in A$.
- i and j can be aggregated if $\delta(i,j) \leq \epsilon$.

Mathematically elegant but inefficient:

- Limited by its computational complexity:
  - $O(\log \frac{max_{i,j \in S} \delta(i,j)}{p_{target}} 2^{K|S|})$.

Proposed algorithms: The $\phi_{Q_d^*}$ K-MDP algorithm

Previous algorithm has limitations.

Based on a transitive approximate abstraction function $\phi_{Q_d^*}$ [Abel et al., 2018]:

$$\phi_{Q_d^*}(i) = \phi_{Q_d^*}(j) \implies \forall a \left\lceil \frac{Q^*(i,a)}{d} \right\rceil = \left\lceil \frac{Q^*(j,a)}{d} \right\rceil \tag{4}$$

- A set of states belong to the same cluster if they belong to the same bin.
- Binary search on d:
  - As bins size decrease, abstraction is more precise.
- $O(|S| \log(\frac{\text{VMAX}}{p_{target}}))$.

Proposed algorithms: The $\phi_{a_d^*}$ K-MDP algorithm

Our new state abstraction contribution:

- The approximate transitive state abstraction function $\phi_{a_d^*}$ satisfies:

$$\phi_{a_d^*}(i) = \phi_{a_d^*}(j) \implies a_i^* = a_j^* \wedge \left\lceil \frac{V^*(i)}{d} \right\rceil = \left\lceil \frac{V^*(j)}{d} \right\rceil \tag{5}$$

- More restrictive.
  - States must have same optimal action.
- Abstraction not guarantee if K is smaller than the number of actions in the optimal policy.
- $O(|S| \log(\frac{\text{VMAX}}{p_{target}}))$.

Proposed algorithms: The k-means++ K-MDP algorithm

- State abstraction using a clustering technique.
- Interesting, but we don't have a loss of performance.
- Minimize the average squared distance between points in the same cluster.
- k-means++ with the optimal Q function and norm $L1$ to define the state space S_K:

$$\sum_{s \in S} \min_{s_K \in S_K} ||Q^*(s, \cdot) - Q^*(s_K, \cdot)||^2. \tag{6}$$

- Complexity $O(nkdi)$.

Recovering two endangered species

Problem published in 2012 [Chades et al., 2012];

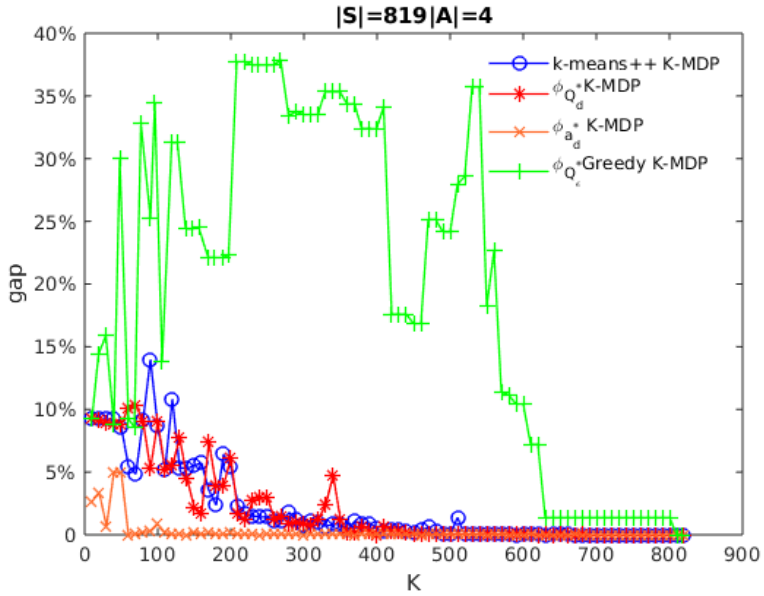
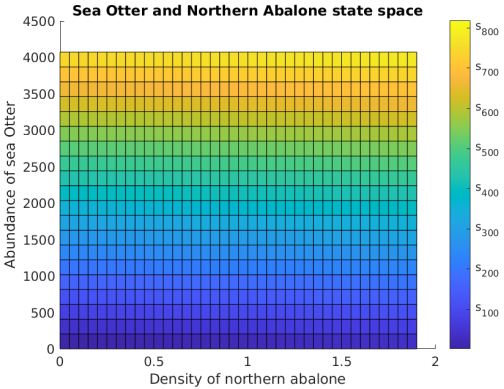
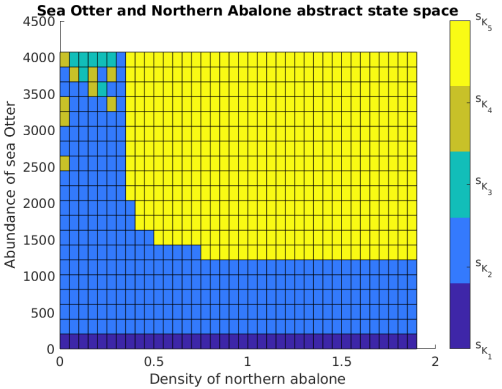


Figure 2: Performance of K-MDP algorithms on the sea otter and northern abalone conservation problem .

Recovering two endangered species: Reduced model



(a) State space representation.



(b) Abstract state space representation.

Figure 3: Sea Otter and Northern Abalone state representation over the abalone density and otter abundance. (a) Original 819 states. (b) K = 5 abstract states.

Recovering two endangered species: Reduced policy

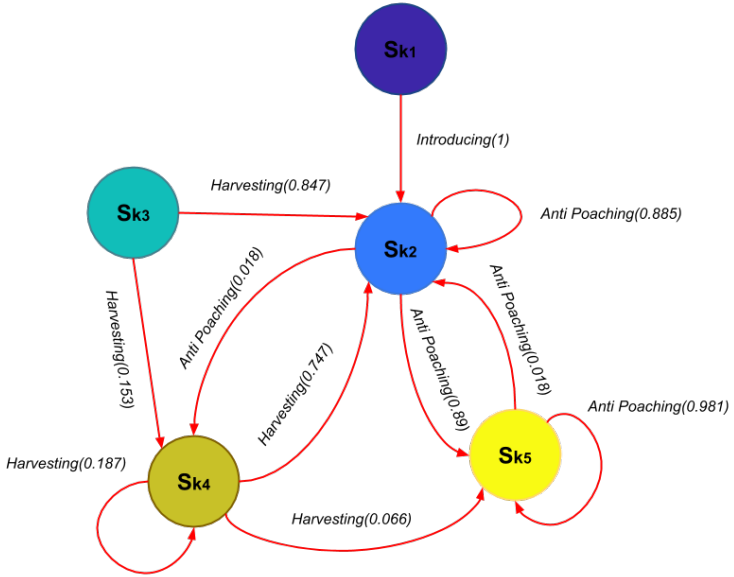


Figure 4: Recovering two endangered species K-MDP model policy graph.

Summary

- Our approach aims at increasing uptake of MDPs in human operated systems by providing easier to interpret models and solutions.
- We have employed transitive approximate state abstraction functions to solve K-MDPs.
- We have proposed to use a clustering technique k-means++ and a binary search greedy approach.

Immediate future work:

- Testing the interpretability of proposed K-MDPs solutions.

Challenges:

- Visualization and interpretability of models and solutions with a large number of state variables.

References

[Abel et al., 2018] Abel, D., Arumugam, D., Lehnert, L., and Littman, M. (2018). State abstractions for lifelong reinforcement learning. In *ICML*.

[Abel et al., 2016] Abel, D., Hershkowitz, D., and Littman, M. (2016). Near optimal behavior via approximate state abstraction. In *ICML*, volume 48 of *PMLR*, pages 2915–2923, New York, NY, USA.

[Brigham and Dutton, 1983] Brigham, R. C. and Dutton, R. D. (1983). On clique covers and independence numbers of graphs. *Discrete Mathematics*, 44(2):139–144.

[Chades et al., 2012] Chades, I., Curtis, J. M. R., and Martin, T. G. (2012). Setting realistic recovery targets for two interacting endangered species, sea otter and northern abalone. *Conservation Biology*, 26(6):1016–1025.

[Dean and Givan, 1997] Dean, T. and Givan, R. (1997). Model minimization in Markov decision processes. In *AAAI/IAAI*, pages 106–111.

[Li et al., 2006] Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a unified theory of state abstraction for MDPs. In *ISAIM*.

[Péron et al., 2017] Péron, M., Jansen, C. C., Mantyka-Pringle, C., Nicol, S., Schellhorn, N. A., Becker, K. H., and Chadès, I. (2017). Selecting simultaneous actions of different durations to optimally manage an ecological network. *Methods in Ecology and Evolution*, 8(10):1332–1341.

Thank You

CSIRO Conservation Decisions Team — Land and Water

Jonathan Ferrer Mestres
t +61 7 3833 5905
e jonathan.ferrermestres@csiro.au
w Conservation Decisions