A Universal 2-state n-action Adaptive Management Solver

Luz Valerie Pascal,[1] **Marianne Akian,**[1] **Sam Nicol,**[2] **Iadine Chades**[2]
[1] INRIA and CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France
[2] CSIRO, Ecosciences Precinct, 41 Boggo Rd, Dutton Park QLD 4102, Australia
luz.pascal96@gmail.com, marianne.akian@inria.fr, sam.nicol@csiro.au, iadine.chades@csiro.au

Abstract

In poor data and urgent decision-making applications, managers need to make decisions without complete knowledge of the system dynamics. In biodiversity conservation, adaptive management (AM) is the principal tool for decision-making under uncertainty. AM can be solved using simplified Mixed Observable Markov Decision Processes called hidden model MDPs (hmMDPs) when the unknown dynamics are assumed stationary. hmMDPs provide optimal policies to AM problems by augmenting the MDP state space with an unobservable state variable representing a finite set of predefined models. A drawback in formalising an AM problem is that experts are often solicited to provide this predefined set of models by specifying the transition matrices. Expert elicitation is a challenging and time-consuming process that is prone to biases, and a strong assumption of hmMDPs is that the true transition matrix will be included in the candidate model set. We propose an original approach to build a hmMDP with a universal set of predefined models that is capable of solving any 2-state n-action AM problem. Our approach uses properties of the transition matrices to build the model set and is independent of expert input, removing the potential for expert error in the optimal solution. We provide analytical formulations to derive the minimum set of models to include into an hmMDP to solve any AM problems with 2 states and n actions. We assess our universal AM algorithm on two species conservation case studies from Australia and randomly generated problems.

Introduction

In many computational sustainability domains such as conservation of biodiversity, epidemiology or natural resource management, managers must adapt their decisions to the state of the systems and account for future events. When the dynamics are Markovian and the state-transition matrices are known, Markov decision processes (Bellman 1957) are a suitable model to help managers making sequential-decisions under uncertainty (Marescot et al. 2013). Unfortunately, for some of the most pressing problems in conservation, health and biosecurity, data is scarce and transition matrices are rarely available (Chadès and Nicol 2016). In such situations, Adaptive Management (AM) or learning while doing is the recommended management practice (Walters and Hilborn 1978; Keith et al. 2011). Decisions are selected to achieve a management objective while simultaneously gaining information to improve future management success (Holling 1978; McCarthy, Armstrong, and Runge 2012; Walters 1986).

Several approaches have been developed to solve AM problems (Chadès et al. 2017). Here, we focus on the class of problems characterised by model uncertainty. AM under model uncertainty assumes that the unknown true dynamics of the system are close enough to a set of pre-defined models that specify how the system dynamics function. AM tools to reduce model uncertainty were first proposed in the fisheries literature (Silvert 1978).

The key prerequisite for an optimal AM system designed to reduce model uncertainty is that plausible alternative hypotheses about system function dynamics can be articulated. Chadès et al. (2012) have shown that model uncertainty AM problems can be cast as Mixed Observability MDPs (MOMDPs), and under stationary assumptions as hidden model MDPs (hmMDPs). This approach was used to inform conservation of migratory shorebirds under sea level rise (Nicol et al. 2013, 2015) and is now being tested by the New South Wales Saving our Species Program in Australia (Potoroo case study in this paper).

A drawback of model uncertainty approaches is the assumption that the real model is contained within the model set. As well as making this strong assumption, current methods often rely on asking experts to parameterize the model set, which is both time consuming and prone to human error and biases (Martin et al. 2012). Point-based POMDP solvers help us overcome the technical challenge of solving the hmMDP for a large set of models so that we have a greater likelihood of including the real model. However, in data-poor systems, including many models makes it difficult to distinguish the real model as observations are few. Problems in AM application domains often have limited observations. In these situations, it is critical to only include a small model set. We postulate the existence of a parsimonious set of models that spans the set of possible optimal solutions for AM problems using as few models as possible. Finding this minimum, yet universal set of models is the objective of our manuscript. We will focus on 2-state problems, where we rely on statements of relativity (i.e. "good" and "bad" states) to generate universal output.

Our approach explores the parameter space of the un-

known state-transition matrices to derive a minimum but empirically robust set of representative models for a 2-state n-action universal AM solver. We demonstrate the usefulness of our approach on AM conservation problems of two listed threatened species (Gouldian finch and Long-footed Potoroo) and randomly generated problems. The code and data is freely available at https://github.com/conservation-decisions/Universal-Adaptive-Management-Solver.

We first introduce some related work and the necessary background. We explain how AM problems can be modelled as hidden Markov MDPs. We then define our problem and propose a new algorithm MC-UAMS that takes advantage of key properties and propositions to build a hmMDP for 2-state n-action problem. Finally, we assess our approach and discuss future research.

Related Work

In the Bayesian Reinforcement Learning (BRL) setting, agents try to maximize the collected rewards while interacting with their environment while using some prior knowledge that is accessed beforehand. Many BRL algorithms have already been proposed (Duff 2003; Dearden, Friedman, and Russell 1998; Poupart et al. 2006). The most relevant approach is described by Dearden, Friedman, and Andre in 1999. The authors model the unknown parameters using a Dirichlet distribution which is updated by counting the number of successes and failures. Poupart et al. (2006) propose to parameterize the optimal value function by a set of multivariate polynomials which are then used to derive an approximate policy optimization algorithm (BEETLE). BEETLE is shown to be efficient and tractable for a small number of unknown parameters, which can grow in $\mathcal{O}(|X|^2|A|)$. Bayesian bandits (Ghavamzadeh et al. 2016) are a useful framework to estimate the values of the outcome probabilities $P(.|a)$. However the framework does not incorporate the influence of the current states and converges in a large number of iterations.

AM problems share some similarities with BRL and Bandits. However, an AM problem is solved as a *planning* problem i.e. the optimal policy is calculated off-line. This is because AM is applied in data-poor contexts which does not allow the extent of system exploration required by traditional reinforcement learning approaches. AM requires thinking ahead and calculating the consequences of all possible values of the unknown information before deciding the optimal action.

MDPs and hmMDPs

Markov Decision Processes (MDPs) are a convenient model to solve sequential decision-making processes under uncertainty (Bellman 1957). Formally, a discrete MDP is defined by a tuple $\langle X, A, T, r, H, \gamma \rangle$. X is a finite set of states. A is a finite set of actions. T is the transition function between states. $T(x, a, x') = P(x_{t+1} = x' | x_t = x, a_H = a)$ represents the probability that the state of the system transitions from x to x' when action a is implemented. r is the reward function. $r(x, a)$ represents the immediate reward received when the current state of the system is x and the implemented action is a. H is the (finite or infinite) horizon. $\gamma \in [0, 1]$ is a discount factor ($\gamma < 1$ if H is infinite). The decision-maker aims to find a sequence of actions that maximizes a selected criterion (Sigaud and Buffet 2010). Here, we will assume that the expected sum of discounted rewards is an appropriate criterion $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0]$. A policy $\pi : X \rightarrow A$ is a mapping from the set of states X to the set of actions A. To a policy π corresponds the value of the criterion, called the value function:

$$V_\pi(x_0) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) | x_0].$$

A policy π is optimal if it maximizes the value function $\pi^*(x_0) = \arg\max_\pi V_\pi(x_0)$. Many exact and approximate algorithms have been designed to solve MDPs, including policy or value iteration (Puterman 1995). One of the main challenges of applying MDPs is the assumption that the transition function is readily available. A solution is to assume that the dynamics of the system is partially observable.

hmMDPs

AM problems assume that the state of the system is completely observable, but the dynamics are uncertain. It is also assumed that the real and unknown MDP belongs to a finite set of predefined models. This problem can be solved using hidden model MDPs (hmMDPs) (Chadès et al. 2012). A hmMDP is a defined as a special case of factored POMDP (also known as as Mixed Observable MDP (Ong et al. 2010)).

A hmMDP is a tuple $\langle X, Y, A, Z, T_x, T_y, O, R, H, b_0, \gamma \rangle$. $X \times Y$ is the factored set of states: X is the set of completely observable variables; Y is the set of partially observable variables. Y is a finite set of hidden models that could represent the dynamics of the system as MDPs. It is assumed that the real model of the dynamics of the system y_r is an element of Y; $O = X$ is the set of observations. We assume that the partially observable variables in Y are not observable; $T_x(x, y, a, x', y') = p(x' | x, y, a)$ gives the probability that the value of the fully observable state is x' at time $t + 1$ when action a is performed in state (x, y) at time t and has already led to y'. We assume the real model y_r will not change over time i.e. T_y is the identity matrix. As only variables in X are observable, the observation function Z is the identity matrix; R is the reward matrix of the immediate reward $R(x, y, a)$ that the policy-maker receives for implementing a in state (x, y); b_0 is an initial belief, a probability distribution over partially observable states; γ, A and H are defined as in the MDP case.

Because the state y is not perfectly observable, it is modeled by a belief state b that is a probability distribution among the elements of Y (Astrom 1965). The set of all belief states is the belief space, denoted B.

A hmMDP policy is defined as $\pi : X \times B \rightarrow A$. A policy π is optimal if it maximizes the selected criterion, $\pi^* = \arg\max_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(x_t, b_t, \pi(x_t, b_t)) | x_0, b_0]$ with $\mathcal{R}(b, x, a) = \sum_{y \in Y} b(y) R(x, y, a)$. Any policy π can be assessed through its value function V_π defined as, for all $x, b \in X \times B$:

$$V_\pi(x,b) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(x_t, b_t, \pi(x_t, b_t))|x, b].$$

We then have $\pi^* = \arg\max_\pi V_\pi(x_0, b_0)$. The optimal value function is denoted V^*. hmMDPs have the same complexity as POMDPs (Chadès et al. 2012), therefore hmMDPs are PSPACE-complete in finite horizon (Papadimitriou and Tsitsiklis 1987), and are undecidable in infinite-horizon (Madani, Hanks, and Condon 1999). Approximate POMDP solvers have been adapted to solve MOMDPs (Ong et al. 2010), and hmMDPs (Péron et al. 2017).

In ecology, a common practice is to build the set Y of predefined models by asking experts to specify a set of possible models (Chadès et al. 2012). The traditional way of solving adaptive management problems under uncertainty relies on the assumption that the true transition matrix is included in the candidate model set. This is a common assumption in applied domains that use this approach such as conservation of biodiversity (Nicol et al. 2013, 2015), natural resource management (Martin et al. 2009) and more recently epidemiology (Shea et al. 2020). Assuming that the real model is one of the experts' models is a questionable assumption. In practice, if the real model y_r is close to one of the models of Y, then the optimal policy of the hmMDP π_Y is still a good policy to solve the problem. The purpose of our manuscript is to build a hmMDP with a *universal* set Y^* of predefined models, meaning that the resulting hmMDP is able to solve any 2-state n-action AM problem.

A 2-state n-action Universal AM Solver

Problem Formulation and Notations

Let us define M as a 2-state $X = \{L, H\}$ and n-action $A = \{a_1, ..., a_n\}$ MDP with unknown stationary state-transition probabilities T and known reward function r. M is the real MDP that decision-makers aim to solve.

The set of MDPs that are potential candidates to be the true MDP is infinite. We define this set as $\mathcal{M}_n^2$. All MDPs in $\mathcal{M}_n^2$ have the same reward function r.

We parameterize $M \in \mathcal{M}_n^2$ in the following way. The transition probabilities depend on $2n$ unknown parameters $(p_L^1, ..., p_H^n)$. For all $k \in \{1, ..., n\}$, $T(a_k) = \begin{pmatrix} p_L^k & 1-p_L^k \\ p_H^k & 1-p_H^k \end{pmatrix}$. $(p_L^1, ..., p_H^n)$ are assumed to follow a known probability distribution μ on $[0,1]^{2n}$. μ is absolutely continuous with respect to Lebesgue's measure (λ). In simple terms this means that there exists a Lebesgue function f_μ that can be integrated (the probability density function) on the real line such that $\mu([0,1]^{2n}) = \int_{[0,1]^{2n}} f_\mu d\lambda$. We write the values of the reward function as $r(s,a) = r_s^a$.

Because most conservation problems aim to maintain or increase a population from "Low" (L) to "High" (H) density, we will assume that for all actions a, b, $r_L^a < r_H^b$. This assumption is true for our two case studies and reasonable for the conservation AM literature e.g. McCarthy and Possingham (2007). We discuss this assumption in the conclusion.

State	$\pi_{1,1}$	$\pi_{1,2}$	...	$\pi_{n,n-1}$	$\pi_{n,n}$
Low	a_1	a_1	...	a_n	a_n
High	a_1	a_2	...	a_{n-1}	a_n

Table 1: Possible optimal policies for a 2-state, n-action problem

We define the probability distribution on $\mathcal{M}_n^2$ as the probability distribution of its parameters $(p_L^1, ..., p_H^n)$. Hereafter, for sake of simplicity we will denote $\mu(\mathcal{M}_n^2) = \mu([0,1]^{2n})$ the volume of $\mathcal{M}_n^2$.

Our problem is to compute a set Y of models in $\mathcal{M}_n^2$ such that our solver is able to find a reliable optimal policy for any real MDP in $\mathcal{M}_n^2$. We call this set Y a *universal* set.

Definition 1. *Let M be an MDP of $\mathcal{M}_n^2$ drawn using the probability distribution μ. Let Y be a finite set of models of $\mathcal{M}_n^2$, and $\mathcal{H}(Y)$ the corresponding hmMDP. Let π_Y^* be the optimal policy of $\mathcal{H}(Y)$. The expected cumulative discounted rewards received when applying π_Y^* to M is denoted V_M^Y. We define $Gap(Y) = \mathbb{E}_\mu[\frac{V_M^* - V_M^Y}{V_M^*}]$ the **expected relative gap to the optimal**.*

$Gap(Y)$ represents the relative error induced by the set of models Y, when managing any model M. Formally, the problem we aim to solve is:

$$Y^* = \arg \min_{\substack{Y \subset \mathcal{M}_n^2 \\ |Y| < \infty}} Gap(Y) \tag{1}$$

To guide the calculation of our universal set of models, we first need to present relevant properties.

Property 1. *Let M be a discrete MDP with stationary transition matrices. Let the expected infinite sum of discounted rewards be the optimization criterion of the problem. Then, the set of possible optimal policies $\pi : X \to A$ is finite, and denoted Π. In particular, for a 2-state n-action problem, there are n^2 possible optimal policies, listed in Table 1.*

Proof. As we are looking for optimal strategies in an infinite discounted horizon and as the dynamics of the system are stationary, the optimal policy is stationary, and can be written $\pi^* : X \to A$. Considering that the set of states X and set of actions A are discrete and finite, there are $|A|^{|S|}$ possible optimal policies. □

Property 1 is important because although the set of MDPs $\mathcal{M}_n^2$ is infinite, the set Π of policies is finite. Using property 1, we can split $\mathcal{M}_n^2$ into distinct spaces of MDPs sharing the same optimal policy.

Definition 2. *We define $\mathcal{M}^\pi$ the subset of MDPs in $\mathcal{M}_n^2$, such that π is an optimal policy.*

$\mathcal{M}^\pi$ is defined by a subset of $[0,1]^{2n}$ denoted $\mathcal{P}^\pi$ that we will define formally in the next section. Similarly, we denote $\mu(\mathcal{M}^\pi) = \mu(\mathcal{P}^\pi)$ the volume of $\mathcal{M}^\pi$ i.e. the probability of π being the optimal policy.

Logically, the set Y of hidden models must contain at least one model per $\mathcal{M}^\pi$ to guarantee that the best action can be chosen. For example, for a 2-state 2-action problem,

Notation	Meaning
$(p_L^1, ..., p_H^n)$	Set of parameters describing the state-transition probabilities for a 2-state n-action MDP
μ	Known probability distribution of the parameters $(p_L^1, ..., p_H^n)$
Π	Finite set of possible optimal policies of a 2-state n-action AM problem
π	Element of Π
$\mathcal{M}_n^2$	Infinite set of 2-state n-action MDPs sharing the same known reward function
$\mathcal{M}^\pi$	Infinite set of 2-state n-action MDPs sharing the same known reward function and having π as optimal policy
$\mathcal{P}^\pi$	Subset of $[0,1]^{2n}$ describing $\mathcal{M}^\pi$
y^π	Universal model of $\mathcal{M}^\pi$. Model minimising the distance with all the other models of $\mathcal{M}^\pi$

Table 2: Notations guide

the set Y must count at least 4 hidden models $y^{\pi_{1,1}}$, $y^{\pi_{1,2}}$, $y^{\pi_{2,1}}$ and $y^{\pi_{2,2}}$ in $\mathcal{M}^{\pi_{1,1}}$, $\mathcal{M}^{\pi_{1,2}}$, $\mathcal{M}^{\pi_{2,1}}$ and $\mathcal{M}^{\pi_{2,2}}$ respectively. While it is tempting to include a large set of hidden models to discretize $\mathcal{M}_n^2$ as precisely as possible, the constraint on the number of observations (data poor) reduces the likelihood of identifying the real model. Besides, the more models we include, the more difficult it is to solve the hmMDP. For these reasons, it is important that few models are contained in the selected set.

Definition 3. *Let $\mathbb{M}$ be the set of $\mathcal{M}^\pi$ such that $\mu(\mathcal{M}^\pi) > 0$. A 2-state n-action **Universal Adaptive Management Solver** (UAMS) is a hmMDP with one hidden model per element of $\mathbb{M}$ such that the UAMS minimizes the expected relative gap to the optimal (eq 1).*

Our challenge is to find the parameters that characterize the best set of hidden models Y^* of a UAMS. To approximate the hidden models of Y^*, we propose to compute for each $\pi \in \Pi$, the model that minimises the distance to all the MDPs in $\mathcal{M}^\pi$, hereafter called the universal model of $\mathcal{M}^\pi$.

Definition 4. *Let $\pi \in \Pi$ such that $P_\mu(\mathcal{M}^\pi) > 0$. We define y^π **the universal model** of $\mathcal{M}^\pi$, with $X = \{L, H\}$, $A = \{a_1, ..., a_n\}$, r_x^a and state-transition probabilities defined with the parameters $({p_L^1}^\pi, ..., {p_H^n}^\pi)$ minimising the distance with all the other MDPs of $\mathcal{M}^\pi$.*

MC-UAMS

We now introduce our Monte Carlo UAMS algorithm (MC-UAMS). MC-UAMS estimates the state-transition probabilities of the universal models of Y^* using Monte Carlo simulations. This algorithm is based on theoretical results that we will develop in the next sections.

MC-UAMS (Alg. 1) requires as input a reward function r and μ a set of distribution functions over the parameters $(p_L^1, ..., p_H^n)$. Without loss of generality we will assume that μ is uniform.

First, the algorithm computes a set Π of possible policies using function *PossiblePolicies(r)*. This function inputs a reward function r, and returns a set Π of policies such that $\forall \pi \in \Pi, \mu(\mathcal{M}^\pi) > 0$. We will see that we can reduce this set by taking advantage of properties of the problem (proposition 2).

Second, function *DrawParameters*(Π, μ, N) aims to gather a large enough representative set of parameters $(p_L^1, ..., p_H^n)$ corresponding to each policy and according to μ. This function randomly draws N sets of parameters $(p_L^1, ..., p_H^n)$. For each draw, it calculates and stores the optimal policy of the corresponding MDP.

Third, for each element π of Π, the function *UniversalModel()* computes the universal model of $\mathcal{M}^\pi$ (line 5). Proposition 3 (next section) states that the state-transition probabilities of the universal model of $\mathcal{M}^\pi$ are defined by the expectation of the parameters $(p_L^1, ..., p_H^n)$ in $\mathcal{P}^\pi$. Function *UniversalModel()* empirically computes this expected value and returns the empirical universal model. The obtained model is added to the current list of models Y.

Finally, the algorithm solves the hmMDP corresponding to the tuple $< X, Y, A, r, H, b_0, \gamma >$ (line 8) by calling point-based MOMDP solvers (Ong et al. 2010).

We note that the complexity of MC-UAMS is the same as the complexity of a hmMDP (Chadès et al. 2012). The complexity of Algorithm 1 for a 2-state n-action problem comes from two functions. Function *DrawParameters* (Line 2) solves at least $K * n^2$ MDPs, with K a constant representing the number of draws for each policy (n^2) to build a reliable approximation of each universal model. In practice, we found that $K = 30$ draws was sufficient to apply the central limit theorem and approximate the set of universal models. A MDP can be solved in a polynomial time, for example using the policy iteration algorithm on a 2-state n-action problem, the optimal solution can be found in a time $O(\frac{n^2}{(1-\gamma)} log(\frac{1}{1-\gamma}))$ (Littman, Dean, and Kaelbling 1995). Therefore, the complexity of the function *DrawParameters* is polynomial in time $O(\frac{n^4}{(1-\gamma)} log(\frac{1}{1-\gamma}))$. MDPs are known to be P-complete (Papadimitriou and Tsitsiklis 1987). Function *Solve_hmMDP* solves the resulting hmMDP. A solution with error epsilon can be found in time $O(|C|^2 + |C| log[\frac{(1-\gamma)\epsilon}{2Rmax}])$ where C is a proper delta cover of the optimal reachable space of the initial belief state (Kurniawati, Hsu, and Lee 2008). The general problem of solving a hmMDP is PSPACE-complete (Chadès et al. 2012).

Algorithm 1 *Monte Carlo UAMS*

Require: $< X, A, r, H, b_0, \gamma >$, μ: distribution of the parameters $(p_L^1, ..., p_H^n)$
1: $\Pi \leftarrow PossiblePolicies(r)$
2: $MatParPol \leftarrow DrawParameters(\Pi, \mu, N)$
3: $Y \leftarrow$list()
4: **for** $\pi \in \Pi$ **do**
5: $\quad y^\pi \leftarrow UniversalModel(\pi, X, A, r, \mu, MatParPol)$
6: $\quad$ add y^π to Y
7: **end for**
8: $\pi^* \leftarrow Solve\_hmMDP(X, Y, A, r, H, b_0, \gamma)$
9: **return** π^*

Our algorithm requires defining two key functions: *PossiblePolicies* and *UniversalModel*. To do so, we will first recall essential properties of the optimal value function of an

MDP. Then, we will use these properties to generate for each π in Π the set of MDPs $\mathcal{M}^\pi$. Finally, we will use the explicit formulations and the properties of $\mathcal{M}^\pi$ to derive Y^*.

Properties of Optimal Value Function and Expected Discounted Cumulative Rewards

The set $\mathcal{M}^\pi$ is defined using the following properties of the expected discounted cumulative rewards.

Property 2. *For a given MDP M in $\mathcal{M}_n^2$, $V_M^\pi(x_0) = \mathbb{E}[\sum_{t=0}^{\infty} r(x_t, \pi(x_t))|x_0]$, the expected discounted cumulative rewards of a policy $\pi : X \to A$, is fully determined by the transition probabilities between states $(p_L^1, ..., p_H^n)$. Let $(i, j) \in \{1, ..., n\}^2$:*

$$\begin{aligned} V^{\pi_{i,j}}(L) &= \frac{r_L^{a_i}(1-\gamma+\gamma p_H^j) + r_H^{a_j}(\gamma - \gamma p_L^i)}{(1-\gamma)(1-p_L^i\gamma+p_H^j\gamma)} \\ V^{\pi_{i,j}}(H) &= \frac{r_H^{a_j}(1-p_L^i\gamma) + r_L^{a_i} p_H^j \gamma}{(1-\gamma)(1-p_L^i\gamma+p_H^j\gamma)} \end{aligned} \tag{2}$$

Proof. (Sketch) To each policy $\pi_{i,j}$, corresponds a state-transition matrix $T^{\pi_{i,j}} = \begin{pmatrix} p_L^i & 1-p_L^i \\ p_H^j & 1-p_H^j \end{pmatrix}$ and a reward matrix $R^{\pi_{i,j}} = \begin{pmatrix} r_L^{a_i} \\ r_H^{a_j} \end{pmatrix}$. Let x_0 be a vector indicating if the initial state is "L" or "H" i.e. if the initial state is "L", $x_0 = [1, 0]$, and $x_0 = [0, 1]$ otherwise. $V^{\pi_{i,j}}(x_0)$ rewrites $V^{\pi_{i,j}}(x_0) = \sum_{t=0}^{\infty} \gamma^t x_0.(T^{\pi_{i,j}})^t.R^{\pi_{i,j}}$ where "." is the inner product. Using the eigenvalues of $T^{\pi_{i,j}}$, one can diagonalize $T^{\pi_{i,j}} = P\Delta P^{-1}$, with P an invertible matrix and Δ a diagonal matrix. Then $(T^{\pi_{i,j}})^t = P\Delta^t P^{-1}$, and $V^{\pi_{i,j}}(x_0) = \sum_{t=0}^{\infty} \gamma^t x_0.P\Delta^t P^{-1}.R^{\pi_{i,j}}$. $V^{\pi_{i,j}}(x_0)$ becomes a geometric sum, and can be rewritten into the presented equations (2).□

Property 3. *The optimal policy π^* of an MDP verifies $V^{\pi^*}(x) \geq V^\pi(x)$ for all $x \in X$ and all $\pi \in \Pi$.*

Proof is given by the definition of an optimal policy (Bellman 1957; Sigaud and Buffet 2010).□

Properties 2 and 3 mathematically define the set $\mathcal{M}^\pi$ using the parameters characterising the state-transition probabilities of a model $(p_L^1, ..., p_H^n)$.

Mathematical Characterization of the Set $\mathcal{M}^\pi$

Recall that we aim to explicitly define the MDP set $\mathcal{M}^\pi$ to derive the best set of hidden models Y^* that builds our UAMS. As a consequence of properties 2 and 3, for a given $\pi \in \Pi$ we can characterize $\mathcal{M}^\pi$ as:

$$\mathcal{M}^\pi = \left\{ \begin{array}{c} m \in \mathcal{M} \\ \text{s.t. } V_m^\pi(x) \geq V_m^{\pi'}(x) \\ \forall x \in X \text{ and } \forall \pi' \in \Pi \end{array} \right\}. \tag{3}$$

This definition states that $\mathcal{M}^\pi$ is an infinite set of MDPs that share the same optimal policy π. Despite $\mathcal{M}^\pi$ being infinite, we will exploit this formulation to show that it is defined by a finite set of equations. This definition can be rewritten using the parameters $(p_L^1, ..., p_H^n)$ that define the transition probabilities as follows:

$$\mathcal{M}^\pi = \left\{ \begin{array}{c} m \in \mathcal{M} \\ \text{s.t. } (p_L^1, ..., p_H^n) \in \mathcal{P}^\pi \end{array} \right\} \tag{4}$$

with :

$$\mathcal{P}^\pi = \left\{ \begin{array}{c} (p_L^1, ..., p_H^n) \in [0, 1]^{2n}, \\ \text{s.t. } V^\pi(p_L^1, ..., p_H^n, x) \geq V^{\pi'}(p_L^1, ..., p_H^n, x) \\ \forall x \in X \text{ and } \forall \pi' \in \Pi \end{array} \right\} \tag{5}$$

Intuitively, the sets $\mathcal{P}^\pi$ divide $[0, 1]^{2n}$ into subsets, that guarantee that π is the optimal policy of the corresponding MDP. The next section presents some properties of sets $\mathcal{P}^\pi$, that we use to derive the best parameters of the set of hidden models Y^*. We will show that depending on the values of the rewards r, $\mathcal{P}^\pi$ can be empty.

Properties of Sets $\mathcal{P}^\pi$

Proposition 1. *Let π be in Π. The set $\mathcal{P}^\pi$ is convex and fully defined by $2(n-1)$ linear relationships between the parameters $(p_L^1, ..., p_H^n)$. Let $(i, j) \in \{1, ..., n\}^2$,*

$$\mathcal{P}^{\pi_{i,j}} = \left\{ \begin{array}{c} (p_L^1, ..., p_H^n) \in [0, 1]^{2n}, \\ \textit{s.t. } \forall k \in \{1, ..., n\}, k \neq i, k \neq j : \\ p_L^k > \frac{r_L^{a_k}(1-p_L^i\gamma+p_H^j\gamma) - r_L^{a_i}(1+p_H^j\gamma) + r_H^{a_j} p_L^i \gamma}{(r_H^{a_j} - r_L^{a_i})\gamma} \\ p_H^k > \frac{r_H^{a_k}(1-p_L^i\gamma+p_H^j\gamma) - r_H^{a_j}(1-p_L^i\gamma) - r_L^{a_i} p_H^j \gamma}{(r_H^{a_j} - r_L^{a_i})\gamma} \end{array} \right\} \tag{6}$$

Proof.(Sketch) Using property 2 and the definition of $\mathcal{P}^{\pi_{i,j}}$ in equation 5, we derive equations 6. $\mathcal{P}^{\pi_{i,j}}$ is convex because it intersects convex sets. Note that we used strict inequalities to avoid having sets with $\mu(\mathcal{P}^{\pi_{i,j}}) = 0$.□

Proposition 2. *Recall that $r_L^a < r_H^b$. Linear relationships between the values of the reward function r predict the emptiness of the sets of parameters $\mathcal{P}^\pi$, and thus of the sets of MDPs $\mathcal{M}^\pi$. For given $(i, j) \in \{1...n\}^2$, $\pi_{i,j} \in \Pi$, $\mathcal{P}^{\pi_{i,j}}$ is not empty if for all couples $(k, l) \in \{1...n\}^2, k \neq i, l \neq j$ one of the relationships in table 3 is verified.*

Proof. (Sketch) For a given $\mathcal{P}^\pi$, we explored the conditions on the values of the reward function such that equations defining $\mathcal{P}^\pi$ have a non empty solution set, and are compatible with each other. □

Id.	Relationship
$A_{k,l}$	$r_H^{a_j} \geq r_H^{a_l}$ and $r_L^{a_i} > r_L^{a_k}$
$B_{k,l}$	$r_H^{a_j} \geq r_H^{a_l}, r_L^{a_i} \leq r_L^{a_k}$ and $r_L^{a_k} \leq (1-\gamma) r_L^{a_i} + r_H^{a_j}\gamma$
$C_{k,l}$	$r_L^{a_i} > r_L^{a_k}, r_H^{a_j} < r_H^{a_l}$ and $r_H^{a_j} \geq r_H^{a_l}(1-\gamma) + \gamma r_L^{a_i}$
$D_{k,l}$	$r_H^{a_j} < r_H^{a_l}, r_L^{a_i} \leq r_L^{a_k}$ $r_L^{a_k} \leq (1-\gamma) r_L^{a_i} + r_H^{a_j}\gamma$ and $r_H^{a_j} \geq r_H^{a_l}(1-\gamma) + \gamma r_L^{a_i}$

Table 3: Linear relationships between rewards ensuring that $\mathcal{P}^{\pi_{i,j}}$ is not empty.

Building on proposition 1, proposition 2 shows that we can reduce the number of possible optimal policies in Π given some linear relationships between values of the reward function by investigating conditions under which a given $\mathcal{P}^\pi$ is not empty. In practice, this proposition will speed-up the derivation of the UAMS (function $DrawParameters$ line 2 of Alg.1).

Our next challenge is to derive the transition probabilities of the hidden models of the set Y^*.

Universal Model of $\mathcal{M}^\pi$

Proposition 3. *Let π be in Π such that $\mu(\mathcal{P}^\pi) > 0$. The state-transition probabilities of the universal model y^π are parametrized by the expectancy of the parameters of $\mathcal{P}^\pi$. Formally,*

$(p_L^1{}^\pi, ..., p_H^n{}^\pi) = \mathbb{E}[(p_L^1, ..., p_H^n) | (p_L^1, ..., p_H^n) \in \mathcal{P}^\pi]$.

Proof.(Sketch) We solve the optimization problem

$$\min_{P=(p_L^1,...,p_H^n)\in\mathcal{P}^\pi} \int_{x\in\mathcal{P}^\pi} (P-x).(P-x)^T d\mu(x)$$

$$= \min_{P\in\mathcal{P}^\pi} \int_{x\in\mathcal{P}^\pi} (p_L^1 - x_L^1)^2 + ... + (p_H^n - x_H^n)^2 d\mu(x)$$

The solution $P^* = (p_L^1{}^*, ..., p_H^n{}^*)$ obtained without constraints is: $P^* = \mathbb{E}[P|P \in \mathcal{P}^\pi]$. P^* is the center of mass of $\mathcal{P}^\pi$. Using proposition 1, we know that $\mathcal{P}^\pi$ is convex. Therefore, P^* the center of mass of $\mathcal{P}^\pi$ is included in $\mathcal{P}^\pi$. Thus, the solution of the optimization problem is P^*. □

In practice, proposition 3 shows that for a given $\pi \in \Pi$, we can compute the state-transition probabilities of y^π the universal model of $\mathcal{M}^\pi$ using Monte Carlo simulations (line 5 of Alg. 1).

Experiments

Baseline Algorithm and Settings

We compared MC-UAMS with **PUBD** (Parameter Uncertainty Beta Distribution) – a well known algorithm for solving AM problems under parameter uncertainty (Chadès et al. 2017) and Bayesian RL (Dearden, Friedman, and Andre 1999). PUBD has been applied to AM of wildlife harvest, threatened species translocation and conservation of metapopulations (Hauser and Possingham 2008; Runge, Grand, and Mitchell 2013; Southwell et al. 2016). PUBD solves a *planning* MDP problem by assuming that the unknown parameters follow Beta distributions with parameters (α, β). The distributions are updated at each timestep according to the outcomes. While this approach does not require experts' opinion, the number of hyper-states (number of parameters (α,β) needed to fully describe the current state of the system) grows exponentially with the time horizon H $(|X|^2|A|)^H$ (and is therefore limited to short time horizons (see algorithm in supp. material).

We assessed MC-UAMS on different types of problems (2-state n-action problems denoted 2XnA): two conservation problems (Goudian Finch 2X4A, Long-footed potoroo 2X6A) and 5 randomly generated problems (2X2A, 2X4A, 2X6A, 2X10A and 2X100A). For all experiments, we set $\gamma = 0.9$. We assume a uniform b_0 (no priors on the model sets) and a uniform distribution on the parameters $(p_L^1, ..., p_H^n)$ on $[0, 1]^{2n}$ (no priors on the parameters sets).

We allowed MC-UAMS and PUBD to run for one hour before evaluating their performances via simulations. Only for the problem 2X100A, we allowed MC-UAMS to run for 6 hours due to the large number of possible optimal policies, PUBD ran out of memory for an horizon larger than 2. Performances were evaluated using the empirical expected relative difference to the optimal value function $\widehat{Gap}(.) = \frac{1}{|\mathcal{M}|}\sum_{M\in\mathcal{M}} \frac{V_M^* - V_M}{V_M^*}$ and the relative expected difference between instant rewards (inst. diff.) at the final time step (H=50) $\widehat{Diff} = \frac{1}{|\mathcal{M}|}\sum_{M\in\mathcal{M}} \frac{r_H^* - r_H}{r_H^*}$ (r_H^* is the optimal expected instant reward at time H and r_H is the expected instant reward at time H, for MC-UAMS or PUBD).

Unless stated otherwise, for each problem, we randomly selected a set $\mathcal{M}$ of 100 MDPs according to a uniform distribution on the parameters $(p_L^1, ..., p_H^n)$. Those MDPs represent the "real" dynamics of the system. For each MDP $M \in \mathcal{M}$, we computed the optimal value function V_M^* using the R-package MDPtoolbox (Chadès et al. 2014). Then, we simulated 100 trajectories of 50 time-steps of management of the unknown system M, using the policies derived by MC-UAMS and PUBD. We empirically computed V_M^{UAMS} (resp. V_M^{PUBD}) the expected value of the resulting sum of discounted rewards using the MC-UAMS (resp. PUBD).

Algorithms were implemented in R (version 4.0.2) (R Core Team 2020). hmMDPs (line 8 Alg. 1) were solved using MO-SARSOP (Ong et al. 2010) on Cygwin (version 3.1.6(0.340/5/3), 64 bits). MO-SARSOP is implemented in C++, and was compiled with g++(GCC) 9.3.0. We ran all the experiments on a 230GHz Intel PENTIUM CPU 3550M, and 4Go of RAM. Code, problems and results available at https://github.com/conservation-decisions/Universal-Adaptive-Management-Solver.

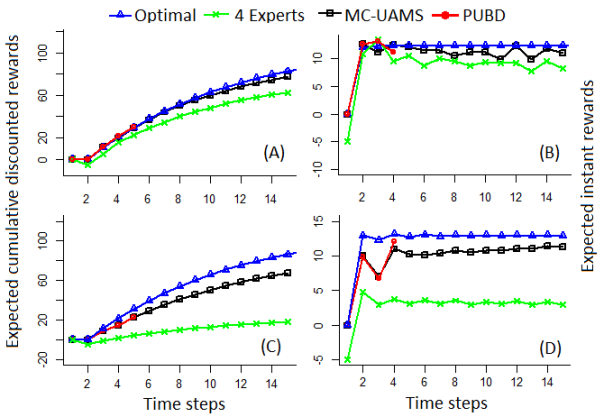

Figure 1: MC-UAMS beats the original Gouldian Finch expert derived hmMDP (4 Experts) in all scenarios. For A and B, Expert 3 is the true model. For C and D, the true model is randomly drawn.

Conservation Problems

2X4A, Gouldian finch: This problem is the conservation problem studied in (Chadès et al. 2012). The aim of the decision makers is to maximize the likelihood of persistence of a population of Gouldian Finch birds in the Kimberley, Australia. The population status (low or high persistence) can be observed, but the response of the population to management actions is unknown. This problem is modeled as an AM problem with 2 states and 4 actions (Do nothing, improve fire and grazing management, control feral cats, provide nesting boxes). Originally, four experts in conservation biology provided four plausible models of the response of the population to management actions. Chadès et al. (2012) solved this AM problem by building a hmMDP with $X = \{Low, High\}$ and $Y_{4Exp} = \{Exp1, Exp2, Exp3, Exp4\}$. We compared the results obtained with the hmMDP built with the 4 experts against our MC-UAMS and PUBD (Table 4).

Our first evaluation **2X4A, Gouldian finch (EM)** assumes the true model is effectively one of the experts (i.e. $Exp3$). Interestingly, both MC-UAMS and PUBD outperformed the 4-experts hmMDP for H=5 (timeout of PUBD) (Figure 1 A and B). For horizons greater than 5, MC-UAMS outperformed the 4-experts hmMDP for both performance criteria ($\widehat{Gap}$= 4% diff. inst.= 10%). The poor performance of the 4-experts hmMDP is not surprising because experts 1 and 2 favoured the same strategy (Fire and Grazing) while Expert 3 (true model tested here) favoured providing nesting boxes with an overall pessimistic outcome for the Gouldian Finch. The policy of the 4-experts hmMDP takes advantage of this information and attempts implementing nesting boxes as a last resort. This is in contrast with MC-UAMS that makes no assumptions on the performance of each management action. MC-UAMS increases its performance over time.

Our second evaluation **2X4A, Gouldian finch** assumes that the true model is uniformly drawn. PUBD and MC-UAMS outperformed the 4-experts hmMDP for H=5 (timeout of PUBD) (Figure 1 C and D) and MC-UAMS largely outperformed the 4-experts hmMDP for horizon greater than 5. Note that the expected MC-UAMS required 16 models, and PUBD required 1910 hyperstates.

2X6A, Long-footed Potoroo: This problem is inspired by a study of the expected impacts of fox predation on the Long-footed Potoroo, a threatened Australian marsupial. We model this problem with 2 states $\{Low, High\}$ and 6 actions that relate to different intensities of baiting to reduce fox numbers. Data for relative cost estimates are obtained from project annual outcome statements for fox control projects in New South Wales. MC-UAMS performed almost as well as PUBD for $H = 4$ ($\widehat{Gap}$= 37%). Overall, MC-UAMS performed consistently for all criteria reported ($34 - 37\%$ away from the optimal). This problem is clearly more difficult to solve than the Gouldian Finch. For this problem, we do not have expert models to compare our results with. This is because for such large number of actions, experts would struggle to provide 36 different models to cover as many alternative models as MC-UAMS. In practice, for this problem we know *a priori* that some actions dominate others (e.g. $p_L^1 > p_L^2$), but our evaluation did not exploit this knowledge. Including this type of prior information into MC-UAMS will likely improve its performance because it will reduce the number of universal models ($|Y^*|$).

	$\widehat{Gap}$ (%) at H (CI 95%)	$\widehat{Gap}$ (%) at ∞ (CI 95%)	diff. inst. (%) at ∞ (CI 95%)
2X4A Gouldian Finch (EM)			
4 Experts hmMDP $\|Y\| = 4$	26 ± 6	28 ± 1	30 ± 15
MC-UAMS $\|Y\| = 16$	$\mathbf{0 \pm 7}$	$\mathbf{4 \pm 2}$	$\mathbf{10 \pm 15}$
PUBD $\|Z\| = 1910$, $H = 5$	5 ± 8	NA	NA
2X4A Gouldian Finch			
4 Experts hmMDP $\|Y\| = 4$	86 ± 7	78 ± 6	75 ± 7
MC-UAMS $\|Y\| = 16$	29 ± 5	$\mathbf{20 \pm 2}$	$\mathbf{6 \pm 1}$
PUBD $\|Z\| = 1910$, $H = 5$	$\mathbf{27 \pm 5}$	NA	NA
2X6A Long-footed Potoroo			
MC-UAMS $\|Y\| = 36$	37 ± 5	$\mathbf{36 \pm 5}$	$\mathbf{34 \pm 5}$
PUBD $\|Z\| = 1324$, $H = 4$	$\mathbf{36 \pm 4}$	NA	NA
2X2A			
MC-UAMS $\|Y\| = 4$	$\mathbf{1.2 \pm 1}$	$\mathbf{3.3 \pm 0.7}$	$\mathbf{1 \pm 1}$
PUBD $\|Z\| = 6614$, $H = 10$	2 ± 5	NA	NA
2X4A			
MC-UAMS $\|Y\| = 12$	6 ± 1	$\mathbf{5 \pm 1}$	$\mathbf{3 \pm 1}$
PUBD $\|Z\| = 6614$, $H = 10$	$\mathbf{6 \pm 1}$	NA	NA
2X6A			
MC-UAMS $\|Y\| = 24$	12 ± 2	$\mathbf{9 \pm 1}$	$\mathbf{4 \pm 1}$
PUBD $\|Z\| = 6614$, $H = 10$	$\mathbf{11 \pm 2}$	NA	NA
2X10A			
MC-UAMS $\|Y\| = 67$	$\mathbf{15 \pm 2}$	$\mathbf{12 \pm 2}$	$\mathbf{7 \pm 2}$
PUBD $\|Z\| = 376$, $H = 3$	$\mathbf{15 \pm 1}$	NA	NA
2X100A			
MC-UAMS $\|Y\| = 574$	$\mathbf{21 \pm 3}$	$\mathbf{18 \pm 1}$	$\mathbf{10 \pm 1}$
PUBD $\|Z\| = 201$, $H = 2$	$\mathbf{21 \pm 3}$	NA	NA

Table 4: Performance evaluation at different time horizon and for different metrics. $|Y|$ is the number of hidden models, H is the horizon reached by PUBD within one hour, $|Z|$ denotes the number of hyperstates included in PUBD for H.

Randomly Generated Problems

To complete the assessment of our algorithms we also generated random problems (2-state n-action problems denoted 2XnA: **2X2A, 2X4A, 2X6A, 2X10A, 2X100A**) for which random reward functions were drawn (see supp. material). Overall, MC-UAMS performed as well or almost as well as PUBD for most problems where PUBD could be run (H=[2,10]). This is an excellent result because MC-UAMS can only update beliefs on Y^* models, while PUBD is effectively updating parameter values of the transition matrices. For infinite time horizon, MC-UAMS performs relatively well with expected gaps ranging from 3% to 35%. while PUBD cannot compute a solution in reasonable time or memory due to the curse of dimensionality (PUBD's state space grows exponentially as the time horizon increases). The study of the rewards matrix (using table 3) enabled to reduce the number of necessary hidden models (from 100 to 67 for **2X10A** and from 100,000 to 574 for **2X100A**).

Conclusion

By proposing a universal adaptive management solver, we are finally addressing a long overdue problematic assumption of existing AM approaches (Chadès et al. 2012) i.e. AM problems are solved assuming that the real model is contained within a predefined model set (Chadès et al. 2017).

In addition to addressing this strong assumption, we are also tackling the reliance on experts to provide plausible model sets, which is a time consuming practice sensitive to human error and biases (Martin et al. 2012). Our new approach and algorithm provide an alternative to existing practices in conservation of biodiversity and natural resource management.

Using the convex properties of the parameter sets, we were able to derive the MC-UAMS algorithm to build a minimum set of models for 2-state n-action AM problems. MC-UAMS performed as well as a common AM approach based on parameter estimation (PUBD; tractable for short horizon problems). Most importantly, MC-UAMS out-performed existing hmMDP with a hidden model set assessed by experts (Gouldian Finch problem).

The assumptions we made to define a MC-UAMS are not restrictive. Importantly, MC-UAMS can also address non-stationarity of the model by changing the transition probabilities of the partially observable states, and detection probabilities can be defined to model the imperfect detection of the states X.

Assuming that for all actions a, b, the values of the reward function verify $r_L^a < r_H^b$ does not change the essence of the properties demonstrated in this paper. Indeed, if this assumption was not verified, the sets of parameters $\mathcal{P}^\pi$ would remain convex, the inequalities defining $\mathcal{P}^\pi$ would be reversed. Simulations would be able to approximate the universal set of models. We have assumed a uniform distribution over the unknown parameters of the transition matrices.

We expect MC-UAMS will benefit from informative priors when available. Our research is the first to provide a general approach to solve universal AM problems, and provides the basis for future research in this area. In particular, there would be value in generalizing our findings to n-state AM problems. In the general case for a p-state n-action problem the equations defining $\mathcal{P}^\pi$ are no longer linear. Under specific assumptions, we might be able to recover the convexity. We hope our research will inspire others to focus on exploring properties of MDPs or POMDPs to derive algorithms that are suitable for urgent decision-making and poor-data problems (Chadès and Nicol 2016).

References

Astrom, K. J. 1965. Optimal control of Markov processes with incomplete state information. *Journal of mathematical analysis and applications* 10(1): 174–205.

Bellman, R. 1957. A Markovian decision process. *Journal of mathematics and mechanics* 679–684.

Chadès, I.; Carwardine, J.; Martin, T. G.; Nicol, S.; Sabbadin, R.; and Buffet, O. 2012. MOMDPs: a solution for modelling adaptive management problems. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Chadès, I.; Chapron, G.; Cros, M.-J.; Garcia, F.; and Sabbadin, R. 2014. MDPtoolbox: a multi-platform toolbox to solve stochastic dynamic programming problems. *Ecography* 37(9): 916–920.

Chadès, I.; and Nicol, S. 2016. Small data call for big ideas. *Nature* 539(31).

Chadès, I.; Nicol, S.; Rout, T. M.; Péron, M.; Dujardin, Y.; Pichancourt, J.-B.; Hastings, A.; and Hauser, C. E. 2017. Optimization methods to solve adaptive management problems. *Theoretical Ecology* 10(1): 1–20.

Dearden, R.; Friedman, N.; and Andre, D. 1999. Model-based Bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI). Morgan Kaufmann*, 150–159.

Dearden, R.; Friedman, N.; and Russell, S. 1998. Bayesian Q-learning. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 761–768.

Duff, M. O. 2003. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes.* Ph.D. thesis, University of Massachusetts Amherst.

Ghavamzadeh, M.; Mannor, S.; Pineau, J.; and Tamar, A. 2016. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436* .

Hauser, C. E.; and Possingham, H. P. 2008. Experimental or precautionary? Adaptive management over a range of time horizons. *Journal of Applied Ecology* 45(1): 72–81.

Holling, C. S. 1978. *Adaptive environmental assessment and management*. John Wiley & Sons.

Keith, D. A.; Martin, T. G.; McDonald-Madden, E.; and Walters, C. 2011. Uncertainty and adaptive management for biodiversity conservation. *Biological Conservation* 144(4): 1175 – 1178. ISSN 0006-3207.

Kurniawati, H.; Hsu, D.; and Lee, W. S. 2008. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Zurich, Switzerland.

Littman, M. L.; Dean, T. L.; and Kaelbling, L. P. 1995. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 394–402.

Madani, O.; Hanks, S.; and Condon, A. 1999. On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *AAAI/IAAI*, 541–548.

Marescot, L.; Chapron, G.; Chades, I.; Fackler, P. L.; Duchamp, C.; Marboutin, E.; and Gimenez, O. 2013. Complex decisions made simple: a primer on stochastic dynamic programming. *Methods in Ecology and Evolution* 4(9): 872–884.

Martin, J.; Runge, M. C.; Nichols, J. D.; Lubow, B. C.; and Kendall, W. L. 2009. Structured decision making as a conceptual framework to identify thresholds for conservation and management. *Ecological Applications* 19(5): 1079–1090.

Martin, T. G.; Burgman, M. A.; Fidler, F.; Kuhnert, P. M.; Low-Choy, S.; McBride, M.; and Mengersen, K. 2012. Eliciting expert knowledge in conservation science. *Conservation Biology* 26(1): 29–38.

McCarthy, M. A.; Armstrong, D. P.; and Runge, M. C. 2012. Adaptive management of reintroduction. *Reintroduction biology: integrating science and management* 12: 256.

McCarthy, M. A.; and Possingham, H. P. 2007. Active adaptive management for conservation. *Conservation Biology* 21(4): 956–963.

Nicol, S.; Buffet, O.; Iwamura, T.; and Chadès, I. 2013. Adaptive management of migratory birds under sea level rise. In *IJCAI-23rd International Joint Conference on Artificial Intelligence-2013*, 2955–2957. AAAI Press.

Nicol, S.; Fuller, R. A.; Iwamura, T.; and Chadès, I. 2015. Adapting environmental management to uncertain but inevitable change. *Proceedings of the Royal Society B: Biological Sciences* 282(1808): 20142984.

Ong, S. C.; Png, S. W.; Hsu, D.; and Lee, W. S. 2010. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research* 29(8): 1053–1068.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1987. The complexity of Markov decision processes. *Mathematics of operations research* 12(3): 441–450.

Péron, M.; Becker, K.; Bartlett, P.; and Chades, I. 2017. Fast-tracking stationary MOMDPs for adaptive management problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Poupart, P.; Vlassis, N.; Hoey, J.; and Regan, K. 2006. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, 697–704.

Puterman, M. L. 1995. Markov decision processes: Discrete stochastic dynamic programming. *Journal of the Operational Research Society* 46(6): 792–792.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Runge, M. C.; Grand, J.; and Mitchell, M. 2013. Structured decision making. *Wildlife management and conservation: contemporary principles and practices. Johns Hopkins Univ. Press, Baltimore, MD* 51–72.

Shea, K.; Runge, M. C.; Pannell, D.; Probert, W. J.; Li, S.-L.; Tildesley, M.; and Ferrari, M. 2020. Harnessing multiple models for outbreak management. *Science* 368(6491): 577–579.

Sigaud, O.; and Buffet, O. 2010. Markov Decision Processes in Artificial Intelligence: MDPs, beyond MDPs and applications. ISTE/Wiley, Hoboken.

Silvert, W. 1978. The price of knowledge: fisheries management as a research tool. *Journal of the Fisheries Board of Canada* 35(2): 208–212.

Southwell, D. M.; Rhodes, J. R.; McDonald-Madden, E.; Nicol, S.; Helmstedt, K. J.; and McCarthy, M. A. 2016. Abiotic and biotic interactions determine whether increased colonization is beneficial or detrimental to metapopulation management. *Theoretical Population Biology* 109: 44–53.

Walters, C. J. 1986. *Adaptive management of renewable resources*. Macmillan Publishers Ltd.

Walters, C. J.; and Hilborn, R. 1978. Ecological optimization and adaptive management. *Annual review of Ecology and Systematics* 9(1): 157–188.