Fast-tracking stationary MOMDPs for adaptive management problems

Martin Peron

Queensland University of Technology

Peter Bartlett,

University of California, Berkeley

Kai Helge Becker (Kai Becker),

University of Strathclyde, Management Science

Iadine Chades,

CSIRO

Fast-tracking Stationary MOMDPs for Adaptive Management Problems

Abstract

Adaptive management is applied in conservation and natural resource management, and consists of making sequential decisions when the transition matrix is uncertain. Informally described as 'learning by doing', this approach aims to trade off between decisions that help achieve the objective and decisions that will yield a better knowledge of the true transition matrix. When the true transition matrix is assumed to be an element of a finite set of possible matrices, solving a mixed observability Markov decision process (MOMDP) leads to an optimal trade-off but is very computationally demanding. Under the assumption (common in adaptive management) that the true transition matrix is stationary, we propose a polynomial-time algorithm to find a lower bound of the value function. In the corners of the domain of the value function (belief space), this lower bound is provably equal to the optimal value function. We also show that under further assumptions, it is a linear approximation of the optimal value function in a neighborhood around the corners. We evaluate the benefits of our approach by using it to initialize the solvers MO-SARSOP and Perseus on a novel computational sustainability problem and a recent adaptive management data challenge. Our approach leads to an improved initial value function and translates into significant computational gains for both solvers.

Introduction

Adaptive management is an approach tailored for achieving a management objective in environmental problems when the system dynamics is partially unknown (Walters and Hilborn 1978), with applications in conservation (Chadès et al. 2012; Runge 2013), fisheries (Frederick and Peterman 1995), natural resource management (Johnson, Kendall, and Dubovsky 2002) and forest management (Moore and Conroy 2006). Over time, we can learn about the system dynamics by analyzing how the system has responded to our actions so far. Some actions might not seem optimal to achieve the management objective given our current knowledge but might be more informative about the system dynamics than others, potentially resulting in better decisions in the future.

The uncertainty about the system dynamics is often modeled by a finite set of scenarios (Walters and Hilborn 1976; Moore and Conroy 2006). Chadès et al. (2012) showed that this problem can be formulated as a mixed observability Markov decision process (MOMDP), a special case of POMDP (partially observable MDP). An optimal MOMDP policy accomplishes the best trade-off between informative and rewarding actions, with regard to a precise management objective (Chadès et al. 2012).

Researchers from other fields have also looked at variations of the same problems: model-based Bayesian reinforcement learning aims to find the best trade-off (Vlassis et al. 2012), but does not assume the transition matrix to belong to a finite given set - instead probabilities are often assumed to follow a Dirichlet distribution (Duff 2003).

In adaptive management, the true transition matrix is commonly assumed to be stationary, i.e. it does not change over time (Walters and Hilborn (1978), Chadès et al. (2012), Runge (2013) to cite a few). We will make this assumption too and will refer to the problem as a stationary MOMDP. Most MOMDP solvers are α-vector-based, i.e. they update a piecewise linear value function converging to the optimal value function (Araya-López et al. 2010; Ong et al. 2010). In practice, the high complexity of stationary MOMDPs (PSPACE-complete; Chadès et al. 2012) leads to very slow convergence for all but trivial problems.

Based on the properties of stationary MOMDPs, we propose an algorithm generating a lower bound of the value function (Proposition 1). We show that it runs in polynomial time (Proposition 2). Any α-vector-based MOMDP solver can be initialized with this lower bound, with a potentially significant reduction of the computation time. Additionally, our lower bound is provably optimal in the corners of the domain of the value function (Proposition 3). Finally, we demonstrate in Theorems 1 and 2 that, under some assumptions, the derivatives of the optimal value function exist and are equal to those of our lower bound in neighborhoods around the corners of the domain, i.e. our lower bound is a linear approximation of the optimal value function.

The paper is organized as follows: we first introduce MOMDPs formally. We then describe our approach to speed up MOMDP solvers. We illustrate the efficiency of our approach on the management of the invasive mosquito Aedes albopictus in an Australian archipelago and on case studies taken from Nicol et al. (2013). The data is freely available at goo.gl/6f4Rh0. In the last section we discuss our approach and the results obtained.

Mixed observability Markov decision process

A partially observable Markov decision process (POMDP) is a mathematical framework to model the impact of sequential decisions on a probabilistic system under imperfect observation of the states (Sigaud and Buffet 2010). MOMDPs are a special case of POMDPs, where the state can be decomposed into a fully observable component and a partially observable component (Ong et al. 2010). Alternatively, they can be seen as MDPs extended with a non-observable component (Fig. 1). MOMDPs can model various decision problems where an agent knows its position but evolves in a partially observable environment, or when the transition matrices or rewards are uncertain. Formally, a MOMDP (Ong et al. 2010) is a tuple $\langle X, Y, A, O, T_x, T_y, Z, R, \gamma \rangle$ in which:

- The state space is of the form $X \times Y$. The current state (x, y) fully specifies the system at every time step. The component $x \in X$ is assumed fully observable and $y \in Y$ is partially observable;
- A is the finite action space;
- $T_x(x, y, a, x') = p(x'|x, y, a)$ is the probability of transitioning from the state (x, y) to x' when a is implemented. $T_y(x, y, a, x', y') = p(y'|x, y, a, x')$ is the probability of transitioning from y to y' when a is implemented and the observed component transitions from x to x'. The process respects the Markov property in that these probabilities do not depend on past states or actions;
- The reward matrix is the immediate reward $r(x, y, a)$ that the policy-maker receives for implementing a in state (x, y);
- O is the finite observation space;
- $Z(a, x', y', o') = p(o'|a, x', y')$ is the probability of observing $o' \in O$ if the state is (x', y') after action a;
- γ is the discount factor (< 1 in infinite time horizon).

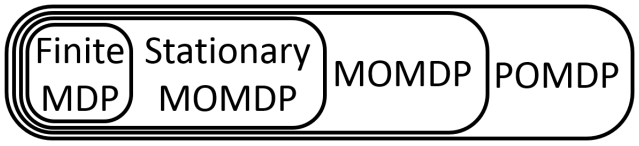

Figure 1: Relations between various Markovian models for sequential decision making.

The sequential decision making process unfolds as follows (Fig. 2a). Starting at time $t = 0$ in a given initial state (x_0, y_0), the decision maker chooses an action a_0 and receives the reward $r(x_0, y_0, a_0)$. The states x_1 and y_1 corresponding to $t = 1$ are drawn according to the probabilities $T_x(x_0, y_0, a_0, \cdot)$ and $T_y(x_0, y_0, a_0, x_1, \cdot)$. The observation o_1 is drawn according to the probability $Z(a_0, x_1, y_1, \cdot)$. The decision maker then observes x_1 and o_1, selects a new action a_1 and the process repeats.

The goal of a decision maker is to find a sequence of actions that yields the best expected sum of rewards over time, depending on the selected criterion. Here, we use an infinite time horizon, i.e. the criterion is $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(x_t, y_t, a_t)|x_0, y_0]$. Because the state y_t is not perfectly observable, it is modeled by a probability vector b_t, called a belief state, where each component represents a state in the set Y (Åström 1965). Belief states are sufficient statistics (Bertsekas 1995), i.e. sufficient knowledge about the system is contained in (x_t, b_t) to make optimal decisions. The set of all belief states is the belief space, denoted B. It is a simplex, i.e. any pair of 'corners' (vertices) are joined by an edge.

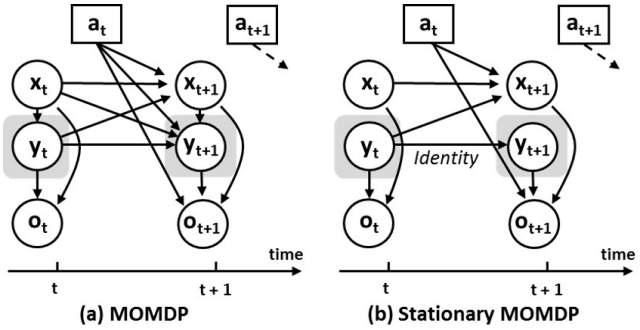

Figure 2: Illustration of the interdependencies between states, observations and actions in a MOMDP and a stationary MOMDP. The grey area surrounding the variable y indicates that it is partially observed.

A MOMDP policy $\pi : X \times B \to A$ is a mapping from the set of components x and belief states b to the set of actions. A policy π is optimal if it maximizes the selected performance criterion:

$$\pi^* = \arg\max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(x_t, b_t, \pi(x_t, b_t))|x_0, b_0] \quad (1)$$

with $R(x, b, a) = \sum_{y \in Y} b(y) r(x, y, a)$. Any policy π can be assessed through its value function V_π defined as, for all $x, b \in X \times B$:

$$V_\pi(x, b) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(x_t, b_t, \pi(x_t, b_t))|x, b], \quad (2)$$

We then have $\pi^* = \arg\max_\pi V_\pi(x_0, b_0)$. Its optimal value function is denoted V^*.

An essential property of POMDPs that translates to MOMDPs is that the value function $V_\pi(x, \cdot)$ is piecewise linear convex (PWLC) in the belief state b for finite horizon problems (Smallwood and Sondik 1973). That is, there exists a finite set Γ_x of $|Y|$-tuples (called α-*vector* hereafter) such that:

$$V_\pi(x, b) = \max_{\alpha \in \Gamma_x} b \cdot \alpha \quad (3)$$

where $b \cdot \alpha = \sum_{y \in Y} b(y)\alpha(y)$ is the inner product. In infinite horizon problems, the value function is only guaranteed to be convex, and can be approximated arbitrarily closely by PWLC functions. Initialized with a lower bound of the optimal value function, most MOMDP solvers calculate the policy by updating the sets Γ_x recursively through Bellman's equation, causing V_π to increase until it is close enough to

the optimal value function. To apply the policy, since each α-vector is associated with an action, the best action to implement at any time step is found by selecting the α-vector that maximizes $b \cdot \alpha$ in Eq. 3. This necessitates knowing the belief state b, which can be calculated recursively. Given the current belief state b_t, the current and future state x and x', the action a and the future observation o', the future belief state b_{t+1} is unique and calculated as follows:

$$\begin{aligned} b_{t+1}(y') &= p(y'|x, b_t, a, x', o') \\ &= \frac{p(o'|x, b_t, a, x', y')p(y'|x, b_t, a, x')}{p(o'|x, b_t, a, x')} \\ &= \eta Z(a, x', y', o') \times \\ &\quad \sum_{y \in Y} T_x(x, y, a, x') T_y(x, y, a, x', y') b_t(y) \end{aligned} \tag{4}$$

where $\eta = 1/p(o', x'|x, b_t, a)$ is a normalizing term.

Stationary MOMDPs

We call a MOMDP 'stationary' when its partially observable component y is stationary, i.e. it will not change over time. Potential examples include a customer's profile or a patient's condition, which can be reasonably assumed stationary over a short period of time. Regarding adaptive management problems, the partially observable component y represents the transition matrix, while the component x models the observed 'physical' system (Chadès et al. 2012). The transition matrix is typically assumed stationary (Walters and Hilborn 1978; Chadès et al. 2012; Runge 2013). This means $T_y(x, y, a, x', y') = 1$ if $y = y'$, 0 otherwise (Fig. 2b). In this case, the future belief state can be written (Chadès et al. 2012):

$$b_{t+1}(y') = \eta Z(a, x', y', o') T_x(x, y', a, x') b_t(y') \tag{5}$$

Proposed approach

In this section we describe how the structure of a stationary MOMDP can be exploited to speed up any α-vector-based MOMDP solver.

Property of stationary MOMDPs

Assume that, at a certain time step t, the transition matrix is known, i.e. $b_t(y) = 1$ for some $y \in Y$ and $b_t(\widetilde{y}) = 0$ for all $\widetilde{y} \neq y$. This belief state is a corner of the belief space B and is denoted by the unit vector e_y. Note that the belief space is a simplex, so by 'corner' we refer to its vertices.

In a stationary MOMDP, the corner e_y is absorbing (i.e. $b_{t'} = e_y$ for all $t' \geq t$), since for all $\widetilde{y} \neq y$, $b_{t+1}(\widetilde{y}) = \eta Z(a, x', \widetilde{y}, o') T_x(x, \widetilde{y}, a, x') \times 0 = 0$, so $b_{t+1}(y) = 1$ (the observable component x may still change). From time step t on, the process is a fully observable Markov decision process, with state space X, action space A, transition matrix $T_{x|y}$ and rewards $r_{|y}$. The new transition matrix and rewards are the restriction of the MOMDP components to the state y: $T_{x|y}(x, a, x') = T_x(x, y, a, x')$ and $r_{|y}(x, a) = r(x, y, a)$.

Algorithm

Our approach (Algorithm 1) builds on this property to generate a lower bound of the optimal value function. First, these $|Y|$ MDPs that correspond to the corners of the belief space are solved (line 2), providing $|Y|$ optimal MDP policies π_y^* and values V_y^*. Then, each policy is evaluated on the $|Y| - 1$ other MDPs (line 6). The combination of these evaluations yields, for each policy, one α-vector per state X (line 8). So, there are $|X||Y|$ α-vectors generated in total. The function $Init$ is defined for any $x, b \in X \times B$ as the maximum over these α-vectors.

Algorithm 1 Calculation of the function $Init$

Input: MOMDP $\langle X, Y, A, O, T_x, T_y, Z, R, \gamma \rangle, T_y = Id$

1: **for** $y \in Y$ **do**
2: $\quad V_y^*, \pi_y^* \leftarrow SolveMDP(X, A, T_{x|y}, r_{|y}, \gamma)$
3: $\quad$ **for** $x \in X$ **do**
4: $\quad\quad \alpha_{x,y}(y) \leftarrow V_y^*(x)$
5: $\quad$ **for** $\widetilde{y} \in Y - \{y\}$ **do**
6: $\quad\quad V_{y,\widetilde{y}} \leftarrow PolicyValue(\pi_y^*, X, A, T_{x|\widetilde{y}}, r_{|\widetilde{y}}, \gamma)$
7: $\quad\quad$ **for** $x \in X$ **do**
8: $\quad\quad\quad \alpha_{x,y}(\widetilde{y}) \leftarrow V_{y,\widetilde{y}}(x)$
9: $Init : (x, b) \mapsto \max_{y \in Y} \alpha_{x,y} \cdot b$, $(x, b) \in X \times B$

Theoretical results

Proposition 1. *The function $Init$ is a lower bound of the optimal value function V^*.*

Proof. Let $y \in Y$. By linearity, the linear functions $(x, b) \mapsto \alpha_{x,y} \cdot b$ equal the value functions of the MOMDP policy consisting of implementing the action $\pi_y^*(x)$ in state $x \in X$, with no regard to the observations of y and no belief state calculation. Consequently, these functions are lower bounds of V^*; so is $Init$ by definition. □

Proposition 2. *Algorithm 1 runs in polynomial time in the number of states $|X|, |Y|$ and actions $|A|$.*

Proof. Algorithm 1 consists of solving $|Y|$ MDPs, which can be solved in polynomial time in $|X|$ and $|A|$ (Littman, Dean, and Kaelbling 1995). The evaluation of $|Y|$ MDP policies $|Y| - 1$ times also runs in polynomial time. □

So, the lower bound can be quickly computed and used as an initial value function in any α-vector-based solver. A good initial value function (i.e. not too far from the optimal value function) can be critical for solving large stationary MOMDPs rapidly, since the value function is calculated recursively through Bellman's equation. In the following we show that the lower bound is optimal in all corners e_y:

Proposition 3. *$V_y^*(x) = V^*(x, e_y)$, for all $x, y \in X \times Y$.*

Proof. As discussed above, when $b_t = e_y$ the MOMDP behaves like a classic MDP. Being the optimal MDP value function, V_y^* is by definition no smaller than any other value function, including $V^*(., e_y)$. Conversely, since the process is also part of the MOMDP, the optimal MOMDP function V^* satisfies $V^*(x, e_y) \geq V_y^*(x)$ for all $x \in X$. □

However, optimality in the corners does not imply that the lower bound $Init$ will be close to V^* in the center of the belief space. The following property states that $Init$ is a linear approximation of V^* in neighborhoods around the corners of the belief space.

We prove that, under some assumptions (Assumptions 1 and 2 below), the directional derivatives of V^* in the corners exist and equal those of $Init$. Formally, the directional derivative of V^* in (x, e_y) along a vector d is defined as $\nabla_d V^*(x, e_y) = lim_{h\to 0} \frac{V^*(x,e_y+hd)-V^*(x,e_y)}{h}$. With Assumption 1, the allowed 'directions' d are along the edges of the belief space (Theorem 1). With Assumptions 1 and 2, all directions are allowed (Theorem 2).

First, satisfying Assumption 1 ensures that the optimal MDP policies are optimal *around* the corners, and not just in the corners. For all $(x, a) \in X \times A$, denote $\pi_{x,a}$ the policy selecting a in state x and following π^*_y in other states.

Assumption 1: There exists $y \in Y$ such that, for each $(x, a) \in X \times A$, the optimal MDP policy π^*_y satisfies either:

- $V^*_y(x, e_y) > V_{\pi_{x,a}}(x, e_y)$ (i.e. $\pi^*_y(x)$ strictly better than a in state x);
- Or, for all $\widetilde{y} \in Y$, $T_x(x, \widetilde{y}, \pi^*_y(x), \cdot) = T_x(x, \widetilde{y}, a, \cdot)$ and $r(x, \widetilde{y}, \pi^*_y(x)) = r(x, \widetilde{y}, a)$ (i.e. $\pi^*_y(x)$ and a have identical outcomes in state x).

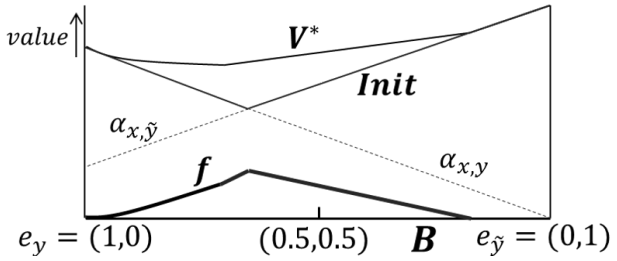

Figure 3: Illustration of an optimal value function V^*, $Init$ and $f = V^* - Init$ between e_y and $e_{\widetilde{y}}$ for a given $x \in X$. In infinite time horizon, the optimal value function V^* is convex but is not necessarily piecewise linear (e.g. near e_y). Under Assumption 1, the derivatives of f in e_y and $e_{\widetilde{y}}$ equals zero, i.e. $Init$ is a linear approximation of V^* in neighborhoods around the corners of the belief space.

In other words, we do not consider cases where for some state x, two optimal actions for transition matrix y have different transition or reward on some transition matrix $\widetilde{y} \in Y$. Under Assumption 1, the directional derivatives of V^* and $Init$ in the corners towards other corners are equal (Fig. 3):

Theorem 1. *We assume that Assumption 1 is satisfied for some $y \in Y$. For all $x \in X$, the directional derivative of the optimal value function in (x, e_y) with respect to any $\widetilde{y} \neq y$ equals that of the function $Init$ (obtained with Algorithm 1). Let $d = e_{\widetilde{y}} - e_y$. For all $x \in X$ and $\widetilde{y} \in Y$, we have:*

$$\nabla_d V^*(x, e_y) = \nabla_d Init(x, e_y) = \alpha_{x,y} \cdot e_y - \alpha_{x,y} \cdot e_{\widetilde{y}} \quad (6)$$

Sketch of proof. (Full proof available in Appendix)

(a) Assumption 1 implies that the optimal MDP policy π^*_y is identical to the optimal MOMDP policy π^* in a neighborhood of the corner (x, e_y).

(b) We show that the belief in transition matrix $\widetilde{y}$ does not grow by more than a constant from one belief state b_t to its successors. The constant equals $\max\{\frac{Z(a,x',\widetilde{y},o')T_x(x,\widetilde{y},a,x')}{Z(a,x',y,o')T_x(x,y,a,x')} | x, x' \in X, a \in A, o' \in O, Z(a, x', y, o')T_x(x, y, a, x') \neq 0\}$.

(c) Combining (a) and (b) applied recursively, we deduce that π^*_y and π^* will be identical for as many time steps as we want, provided b_t is close enough to e_y.

(d) This implies that the distributions of rewards and belief states for π^*_y and π^* will be identical for as many time steps as we want. So, the difference between V^* and $Init$ will only be due to events happening after a number of time steps t' which increases when b_t converges to e_y.

(e) The impact of these future events on $V^* - Init$ can be bounded by $\gamma^{t'} C \|b_t - e_y\|_1$ (with C a constant), which implies that the difference $V^* - Init$ has derivative zero. □

Another assumption on the transition matrices can yield a stronger version of the theorem:

Assumption 2: There exists $y \in Y$ such that, for each $(x, x') \in X$, if $Z(\pi^*_y(x), x', y, o')T_x(x, y, \pi^*_y(x), x') = 0$, then $Z(\pi^*_y(x), x', \widetilde{y}, o')T_x(x, \widetilde{y}, \pi^*_y(x), x') = 0$ for all $\widetilde{y} \in Y$.

In other words, an event that is impossible to observe for transition matrix y cannot be observed for any other transition matrix. This happens, for example, when all scenarios concur on which events are possible and which are not.

Theorem 2. *We assume that Assumptions 1 and 2 are satisfied for some $y \in Y$. Then, for all $x \in X$, the directional derivative of the optimal value function in (x, e_y) in any direction equals that of the function $Init$. For all $(x, b) \in X \times B$, denoting $d = b - e_y$, we have:*

$$\nabla_d V^*(x, e_y) = \nabla_d Init(x, e_y) = \alpha_{x,y} \cdot e_y - b \cdot \alpha_{x,y} \quad (7)$$

So, under Assumption 1, the lower bound $Init$ has the same derivative in the corners as the optimal value function along the edges. Under Assumptions 1 and 2, their directional derivatives in the corners are equal along any direction inside the belief space. These theorems states that the lower bound is a linear approximation of the optimal value function in neighborhoods of the corners of the belief space. We now introduce the real-world case study used to evaluate the validity of our approach.

Case study: managing invasive *Aedes albopictus*

The Asian tiger mosquito *Aedes albopictus* is a known vector of several pathogens. Although the Australian mainland is currently not infested, the nearby Torres Strait Islands are (Ritchie et al. 2006). The $N = 17$ inhabited islands constitute potential sources for the introduction of *Aedes albopictus* into mainland Australia through numerous human-related pathways between the islands and towards north-east Australia (see map in Fig. 4).

Management actions on islands include the treatment of containers and mosquitoes with diverse insecticides. Since

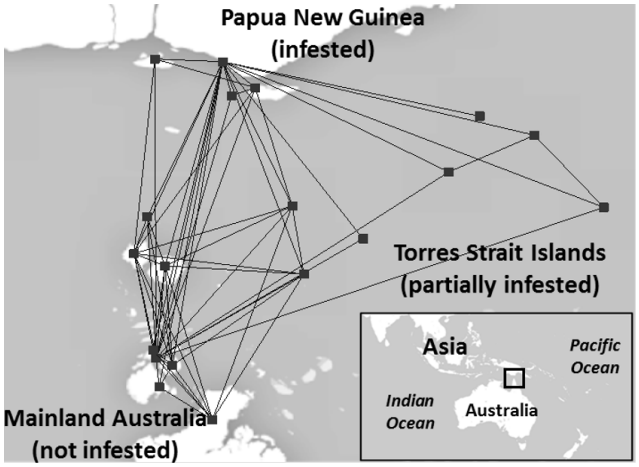

Figure 4: The Torres Strait Islands. Connections between islands depict the possibilities of colonization of the mosquitoes on susceptible islands.

budget is limited, not all islands can be treated simultaneously. The objective is to select islands to manage to maximize the expected time before the mainland becomes infested. The effect of distances and populations on the probability of dispersal between islands and the effectiveness of some of the management actions are partially unknown (i.e. the transition matrix is unknown). A mix of expert data and literature review led us to narrow down the number of transitions matrices to eight. As traditionally in adaptive management, these transitions matrices are assumed equally likely at $t = 0$, i.e. the initial belief state equals $(1/8, \ldots, 1/8)$. This decision problem is modeled as a MOMDP in which:

- The observable component $x \in X$ specifies the season (wet/dry) and the presence or absence of the mosquitoes across the islands N ($|X| = 2^{N+1} + 1$). The last '+1' is an absorbing state corresponding to the presence of mosquitoes in the mainland. The component $y \in Y$ is the unknown true transition function, with $|Y| = 8$;
- Each action $a \in A$ describes which islands should be managed (up to three simultaneously) and the type of management (light or strong);
- The transition probabilities $T_x(x, y, a, x')$ accounts for the possible eradications and transmissions between islands. Also, $T_y(x, y, a, x', y') = 1_{y=y'}$ (stationary);
- The reward $r(x, y, a)$ equals 0 if the mainland is infested and 0.5 otherwise (it only depends on x);
- $O = X$ is the finite observation space;
- $Z(a, x', y', o') = 1_{o'=x'}$ (x' fully observable);
- γ should ideally be 1 so the MOMDP value equals the expected time before infestation of Australia (in years since each time step equals six months). Since most solvers do not support such a setting, we set $\gamma = 0.999$.

We also tested our approach on adaptive management problems of migratory shorebirds taken from Nicol et al. (2013), where we have changed the transition matrix from non-stationary to stationary. We programmed our approach with the MOMDP solver MO-SARSOP (Kurniawati, Hsu, and Lee 2008; Ong et al. 2010) with the MDPSolve package (https://sites.google.com/site/mdpsolve/) and POMDP solver Perseus with 500 beliefs states (Spaan and Vlassis 2005). We compare the modified solvers (marked with a '$^+$') with the original solvers. Note that MO-SARSOP has an advanced lower bound implementation, which we have replaced with our lower bound. MO-SARSOP also initializes an upper bound (fast-informed bound), which is optimal in the corners for all case studies. Perseus initializes its value function with the constant $\frac{1}{1-\gamma} \min_{x,y,a} r(x, y, a)$.

Results

We show the computation times of mosquito instances with number of islands ranging from 7 to 9 (Table 1). Problems for more than 9 islands were not tractable. We show problems *Grey-tailed tattler*, *Red knot pearsonii* and *Red knot rogersi* from Nicol et al. (2013). For problems *Lesser sand plover*, *Terek sandpiper* and *Bar-tailed godwit m*, the initialization is already optimal in MO-SARSOP. Our computer ran out of memory when solving the problems *Great knot*, *Far eastern curlew* and *Curlew sandpiper*.

Table 1: Initial values and initialization times of original and modified (+) MO-SARSOP and Perseus. These are the values of the initial belief state, of the form $(1/|Y|, \ldots, 1/|Y|)$ in all problems. Experiments conducted on a dual 3.46GHz Intel Xeon X5690 with 96GB of memory.

Instance ($\|X\|/\|Y\|/\|A\|$)	MO-SARSOP	MO-SARSOP$^+$	Perseus	Perseus$^+$
7 islands (257/8/113)	11.7 **159 s**	**16.7** 165 s	0 **0 s**	**16.7** 76 s
8 islands (513/8/157)	12.2 **740 s**	**17.4** 771 s	0 **0 s**	**17.4** **169 s**
9 islands (1025/8/211)	12.5 **3244 s**	**17.4** 3316 s	(intractable)	(intractable)
Grey-tailed tattler (972/3/6)	4987 23 s	**5167** **20 s**	836 **0 s**	**5167** 70 s
Red knot pearsonii (8748/3/8)	**6049** 140 s	**6049** **125 s**	4444 **0 s**	**6049** 874s
Red knot rogersi (8748/3/8)	6906 717 s	**6947** **592 s**	(intractable)	(intractable)

Modified solvers consistently obtain a better initial value than original solvers, with the exception of MO-SARSOP on *Red knot pearsonii* (equal value). Moreover, MO-SARSOP$^+$ initializes roughly as quickly as MO-SARSOP. The initialization in Perseus is much quicker than in Perseus$^+$ but at the cost of a lower initial value (0 in our case study because $\min_{x,y,a} r(x, y, a) = 0$).

Fig. 5 illustrates the evolution of the value over time for the original and modified solvers, and the upper bound as calculated in MO-SARSOP$^+$. The modified solvers consistently outperform the original solvers. In our case study *Aedes albopictus*, MO-SARSOP$^+$ obtains much better initial values than MO-SARSOP (7 islands, Fig. 5a). All solvers converge very slowly, which makes this initial value

all the more critical. For *Red knot rogersi* (Fig. 5b), MO-SARSOP$^+$ initializes more rapidly and with a better value than MO-SARSOP, leading to a rapid reduction of the optimality gap (i.e. difference to the upper bound). Regarding *Red knot pearsonii*, our approach does not improve the initial value, but it significantly accelerates the reduction of the optimality gap (Fig. 5c). This supplements Theorem 1 and 2 in suggesting that the generated α-vectors do not solely yield a good value on the initial belief state but all across the belief space, which allows generating good future α-vectors through Bellman's equations. Finally, for all but small problems, Perseus suffers from a poor initial value and is outperformed by Perseus$^+$ (Fig. 5d).

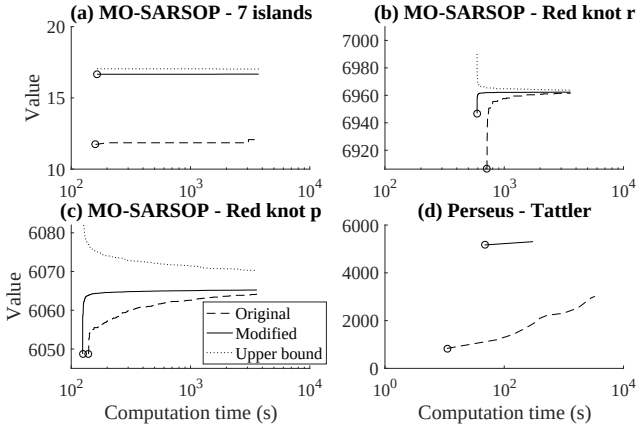

Figure 5: Values over times of original and modified MO-SARSOP and Perseus on 4 problems (stopped after 3600s including initialization). We also show the upper bound as calculated in modified MO-SARSOP. The point corresponding to the initialization time and initial value is circled.

Discussion

We proposed a method to improve the initialization of a MOMDP solver in the case where the partially observable component is stationary. We showed that our approach, which consists of solving a number of Markov decision processes, generates a lower bound that is optimal in the corners of the belief space. With an additional assumption about the optimal policy, we demonstrated that this lower bound is also a linear approximation to the value function. This simple and inexpensive initial lower bound can be used as an initialization to any α-vector-based solver. Tested on two state-of-the-art MOMDP and POMDP solvers, our approach showed significant computational gains on a novel computational sustainability case study of management of an invasive species and on a previously published data challenge.

Our approach has several benefits. It quickly identifies the optimal MDP policies and their values, which solvers may take a very long time to match (Fig. 5a). Since α-vectors are updated recursively through Bellman's equation, α-vector-based solvers very much rely on a good initial value function. Our initial lower bound algorithm has proven to trigger a steeper reduction of the gap in the first steps of computation (Fig. 5b, 5c).

Assumption 1 (two non-identical actions cannot be both optimal) may seem like a strong assumption. However, the set of 'degenerate' instances has measure zero, i.e. a random MOMDP instance will satisfy Assumption 1 with probability 1. As meaningful instances are not random and may well be degenerate, one can slightly perturb their rewards to avoid having two optimal actions. The same goes for Assumption 2, where one can perturb the transition matrices to ensure a transition matrix cannot have probability 0 where other transition matrices have non-zero probability. So, with an arbitrarily small impact on the value of any policy, the assumptions can be fulfilled and the property of linear approximation can be guaranteed.

This property can be exploited in various ways. First, a belief state that is close to a corner can be approximated with the initial value, which would save storage space and backup time. Ideally, the error incurred should be controlled and linked to the distance between the belief state and the corner (also guaranteeing that a policy is near optimal for decision makers), perhaps by bounding the second derivative of the optimal value function. This warrants further research.

The magnitude of the optimality gap after our initialization provides precious information to decision makers. A small optimality gap means that some optimal MDP policies are robust to a transition matrix falsely identified as being true, so adaptive approaches might not be necessary. A large gap shows that a poor knowledge will be heavily penalized and is an incentive to use adaptive methods to reduce the uncertainty; if the value is a financial cost or benefit, this provides an idea of how much money could be spent to reduce uncertainty (value of information; Runge, Converse, and Lyons (2011)).

Our approach could be of use in various contexts of computational sustainability. Stationary MOMDPs are relevant for threatened species management and natural resource management (Runge 2013; Johnson, Kendall, and Dubovsky 2002; Moore and Conroy 2006). In medical science, tradeoffs may occur between learning about a patient's condition and minimising the risk of death, complications, or discomfort (Hauskrecht 1997). In education, an educator may learn a student's profile while teaching in order to identify the best way of teaching (Cassandra 1998).

Apart from computational sustainability, the maintenance of machines, networks or infrastructures (Faddoul et al. 2015) could benefit from our approach, with the partially observable component containing information about the inner state to be maintained, e.g. deterioration or flaws. In marketing, a company or salesperson can learn about the customer as they are implementing their marketing strategy (Zhang and Cooper 2009). Martinelli, Eidsvik, and Hauge (2013) compare diverse approaches to optimize the extraction of oil and gas when the quality of various prospects is uncertain. Dias, Vermunt, and Ramos (2015) infer hidden parameters driving stock markets; Stationary MOMDPs would allow merging the learning and decision processes.

The method can be extended and improved in several ways. Nicol et al. (2013) extended the traditional adaptive

management framework by assuming the transition matrix non-stationary. Our approach does not work under this assumption because in this case the corners of the belief space are not absorbing and so the optimal values on corners cannot be obtained by solving MDPs. However, we hope our research will lead to a stronger focus from the artificial intelligence community on improving lower bounds for general-case MOMDPs or POMDPs. Another common assumption is the finite number of transition matrices; by contrast, Merl et al. (2009) sample continuous parameters with a Monte Carlo approach. We would like to investigate an extension of our current algorithm, to make it compatible with a Monte Carlo approach. Finally, we could not solve very large instances that Nicol et al. (2013) solved with Symbolic Perseus, a factored POMDP solver (Poupart 2005). Our approach could be adapted to factored POMDPs by solving factored MDP (Hoey et al. 1999), also allowing us to solve our case study with a higher number of islands. Decisions problems combining several probabilistic components, such as environmental variability, monitoring efficiency and stakeholder preferences (Convertino et al. 2013), could also benefit from this.

References

Araya-López, M.; Thomas, V.; Buffet, O.; and Charpillet, F. 2010. A closer look at MOMDPs. In *Proc. of the 22nd Int. Conf. on Tools with Artificial Intelligence*, volume 2, 197–204.

Åström, K. J. 1965. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications* 10:174–205.

Bertsekas, D. P. 1995. *Dynamic programming and optimal control*. Athena Scientific, Belmont, MA.

Cassandra, A. R. 1998. A survey of POMDP applications. In *Working Notes of AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes*, 17–24.

Chadès, I.; Carwardine, J.; Martin, T. G.; Nicol, S.; Sabbadin, R.; and Buffet, O. 2012. MOMDPs: A Solution for Modelling Adaptive Management Problems. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Convertino, M.; Foran, C. M.; Keisler, J. M.; Scarlett, L.; LoSchiavo, A.; Kiker, G. A.; and Linkov, I. 2013. Enhanced adaptive management: integrating decision analysis, scenario analysis and environmental modeling for the everglades. *Scientific reports* 3:2922.

Dias, J. G.; Vermunt, J. K.; and Ramos, S. 2015. Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research* 243(3):852–864.

Duff, M. 2003. Design for an optimal probe. In *Proceedings of the 20th International Conference on Machine Learning*, 131–138.

Faddoul, R.; Raphael, W.; Soubra, A.-H.; and Chateauneuf, A. 2015. Partially Observable Markov Decision Processes incorporating epistemic uncertainties. *European Journal of Operational Research* 241(2):391–401.

Frederick, S. W., and Peterman, R. M. 1995. Choosing fisheries harvest policies: when does uncertainty matter? *Canadian Journal of Fisheries and Aquatic Sciences* 52(2):291–306.

Hauskrecht, M. 1997. *Planning and control in stochastic domains with imperfect information*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Hoey, J.; St-Aubin, R.; Hu, A.; and Boutilier, C. 1999. SPUDD: Stochastic planning using decision diagrams. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 279–288. Morgan Kaufmann Publishers Inc.

Johnson, F. A.; Kendall, W. L.; and Dubovsky, J. A. 2002. Conditions and limitations on learning in the adaptive management of mallard harvests. *Wildlife Society Bulletin* 176–185.

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 (RSS)*. 65–72.

Littman, M. L.; Dean, T. L.; and Kaelbling, L. P. 1995. On the complexity of solving Markov decision problems. 394–402. Morgan Kaufmann Publishers Inc.

Martinelli, G.; Eidsvik, J.; and Hauge, R. 2013. Dynamic decision making for graphical models applied to oil exploration. *European Journal of Operational Research* 230(3):688–702.

Merl, D.; Johnson, L. R.; Gramacy, R. B.; and Mangel, M. 2009. A statistical framework for the adaptive management of epidemiological interventions. *PloS One* 4(6):e5807.

Moore, C. T., and Conroy, M. J. 2006. Optimal regeneration planning for old-growth forest: addressing scientific uncertainty in endangered species recovery through adaptive management. *Forest Science* 52(2):155–172.

Nicol, S.; Buffet, O.; Iwamura, T.; and Chades, I. 2013. Adaptive management of migratory birds under sea level rise. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2955–2957. AAAI Press.

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

Poupart, P. 2005. *Exploiting structure to efficiently solve large scale partially observable Markov decision processes*. Ph.D. Dissertation, University of Toronto, Toronto.

Ritchie, S. A.; Moore, P.; Carruthers, M.; Williams, C.; Montgomery, B.; Foley, P.; Ahboo, S.; Van Den Hurk, A. F.; Lindsay, M. D.; and Cooper, B. 2006. Discovery of a widespread infestation of Aedes albopictus in the Torres Strait, Australia. *Journal of the American Mosquito Control Association* 22:358–365.

Runge, M. C.; Converse, S. J.; and Lyons, J. E. 2011. Which uncertainty? Using expert elicitation and expected value of information to design an adaptive program. *Biological Conservation* 144(4):1214–1223.

Runge, M. C. 2013. Active adaptive management for reintroduction of an animal population. *The Journal of Wildlife Management* 77(6):1135–1144.

Sigaud, O., and Buffet, O. 2010. *Markov decision processes in artificial intelligence: MDPs, beyond MDPs and applications*.

Smallwood, R. D., and Sondik, E. J. 1973. Optimal control of partially observable Markov processes over a finite horizon. *Operations Research* 21:1071–1088.

Spaan, M., and Vlassis, N. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research* 24:195–220.

Vlassis, N.; Ghavamzadeh, M.; Mannor, S.; and Poupart, P. 2012. Bayesian reinforcement learning. In *Reinforcement Learning*. Springer. 359–386.

Walters, C. J., and Hilborn, R. 1976. Adaptive control of fishing systems. *Journal of the Fisheries Board of Canada* 33(1):145–159.

Walters, C. J., and Hilborn, R. 1978. Ecological optimization and adaptive management. *Annual Review of Ecology and Systematics* 9:157–188.

Zhang, D., and Cooper, W. L. 2009. Pricing substitutable flights in airline revenue management. *European Journal of Operational Research* 197(3):848–861.