Analysis

Modeling experiential learning: The challenges posed by threshold dynamics for sustainable renewable resource management

Emilie Lindkvist *, Jon Norberg
Stockholm Resilience Centre, Stockholm University, SE-106 91 Stockholm, Sweden

ABSTRACT

Adaptive management incorporates learning-by-doing (LBD) in order to capture learning and knowledge generation processes, crucial for sustainable resource use in the presence of uncertainty and environmental change. By contrast, an optimization approach to management identifies the most efficient exploitation strategy by postulating an absolute understanding of the resource dynamics and its inherent uncertainties. Here, we study the potential and limitations of LBD in achieving optimal management by undertaking an analysis using a simple growth model as a benchmark for evaluating the performance of an agent equipped with a 'state-of-the-art' learning algorithm. The agent possesses no a priori knowledge about the resource dynamics, and learns management solely by resource interaction. We show that for a logistic growth function the agent can achieve 90% efficiency compared to the optimal control solution, whereas when a threshold ( tipping point) is introduced, efficiency drops to 65%. Thus, our study supports the effectiveness of the LBD approach. However, when a threshold is introduced efficiency decreases as experimentation may cause resource collapse. Further, the study proposes that: an appropriate amount of experimentation, high valuation of future stocks (discounting), and, a modest rate of adapting to new knowledge, will likely enhance the effectiveness of LBD as a management strategy.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Exploiting a renewable resource sustainably involves two fundamental constraints: first, the difficulties of agreeing on appropriate actions (Dietz et al., 2003; Hardin, 1968) and second, the limitations in understanding the dynamics of the resource system (Allen et al., 2011; Armitage et al., 2008). This paper will focus on the latter constraint.

1.1. Adaptive Resource Management—The Learning Challenge

Ecosystems are complex adaptive systems (Levin et al., 2012) and thus will always be subject to uncertainty, unknown and unknowable phenomena (Duit and Galaz, 2008; Folke et al., 2005; Levin, 2003). These phenomena include, among others: inherent stochastic events, non-convex interactions, scale dependent dynamics and incomplete information (Crépin et al., 2012; Norberg and Cumming, 2008). Using optimal management (see e.g. Clark, 2010) and predefined models to describe complex resource dynamics with the purpose of deriving an optimal path of action for management or exploitation will always ignore parts of these phenomena and may give a false sense of accuracy in the derived solutions (Allen et al., 2011; Holling and Meffe, 1996; Walters and Hilborn, 1978). The shortcomings of optimization-based 'command and control' management led to the development of adaptive management (Allen et al., 2011; Holling, 1978; Holling and Meffe, 1996). This approach tries to navigate the system by more or less targeted trial and error and by building up a reservoir of knowledge through a continuous learning process (Arrow, 1962; Kolb, 1984; Walters and Holling, 1990). However, adaptive management also faces constraints such as: limitations in building an understanding of the resource dynamics based on iterative and locally based experiences, the cost of experimenting in order to learn about the system and, recognizing and responding to knowledge of changing conditions within the system (Olsson and Folke, 2001).

In this paper we study adaptive management using 'optimal control management' as a benchmark, to explore the limits of learning-by-doing (LBD) when managing a renewable resource exhibiting two levels of non-linear dynamics.

1.2. Thresholds Dynamics

A particularly 'wicked problem' (Jentoft and Chuenpagdee, 2009) in natural resource management is the fact that many ecosystems are subject to threshold dynamics, so called critical transitions or regime shifts (Scheffer, 2009). Threshold effects entail abrupt changes in the dynamics of an ecosystem, where the effect of passing a threshold (or tipping point) switches the dominant feedbacks within the system and can change a particular resource's provisioning rate. Recovery of the resource is then constrained by the degree of lock-in to the new system domain.
i.e., the hysteresis effect (Schefter et al., 2001). Thresholds may be seen as undesirable properties, if the change in provisioning from the system has a significantly negative impact on human well-being (Rockström et al., 2009; Stern, 2007). Moreover, research suggests that regime shifts in human-nature systems are likely to increase as human pressure on the planet accrues (Schlüter et al., 2012).

Depending on the severity of lock-in, thresholds provide a particularly difficult challenge for LBD approaches, should a threshold be crossed while experimenting to learn the dynamics of the system. Recent research is focused on anticipating regime shifts (Biggs et al., 2009; Scheffer et al., 2012), optimal management of systems with regime shifts (Brock and Starrett, 2003; Horan et al., 2011; Polasky et al., 2011), or studying adaptive management in relation to threshold dynamics, using either multi-agent based models for studying emergent properties (Janssen and Carpenter, 1999; Janssen et al., 2000) or laboratory experiments with human subjects (Lindahl et al., 2012). However, the role of threshold dynamics for the LBD processes itself, is in dire need of further exploration.

### 1.3. Introducing Reinforcement Learning and Artificial Intelligence

Decision theories are strongly connected to the learning process, such as the expected utility theory from economics (Neumann and Morgenstern, 1947) and prospect theory from psychology (Kahneman and Tversky, 1979). Reinforcement learning (RL) (Sutton and Barto, 1998) is a computational approach to problems concerning goal-directed learning. It is defined as an agent’s ability to learn a behavior through trial-and-error by interacting with a dynamic environment (Kaelbling et al., 1996), and allows for incorporating different decision theories. Due to its compatibility with adaptive management, RL is perceived as a latent approach for dealing with natural resource management problems (Fonnesbeck, 2005). RL attracts researchers from diverse disciplines such as psychology, control theory, artificial intelligence, and neuroscience (Sutton and Barto, 1998). Notably, Kable and Glimcher (2009) and Niv and Montague (2008) show that it can reveal the neurobiological basis for learning subjective values, which ultimately underlies all decision-making.

RL and neural networks are situated between artificial intelligence and conventional engineering, and “... extend ideas from optimal control theory and stochastic approximation to address the broader and more ambitious goals of artificial intelligence” (Sutton and Barto, 1998). Control theory has contributed to a profound understanding of why complexity in natural systems creates trade-offs between robustness and resilience, and fragility at different scales (Anderies et al., 2007; Cete and Doyle, 2002; Folke, 2006; Levin and Lubchenco, 2008). Furthermore, it provides a well-founded mathematical representation of ‘feedback’ as a process. The RL approach shifts the main focus from control to learning, and accentuates highly theoretical but essential parts of the LBD process—moving it closer to how humans handle information. By combining the component ideas of temporal difference learning (from RL) and neural networks, we incorporate features such as hindsight, planning horizon, exploration vs. exploitation and generalization (further described in Section 2.2).

Learning is a principal aspect of adaptive management in helping to deal with uncertainty and change (Armitage et al., 2008). Social learning has been extensively studied, but the learning process at an individual level has been scarcely addressed in resource management (see e.g. Fazey et al. (2005), Garavito-Bermúdez et al. (in press) and Marschke and Sinclair (2009) for empirical studies). In addition, literature on how the human brain accumulates knowledge through interacting with its environment is not explicitly found in resource management. However, inspiration can be gained from the relatively new field of neuroeconomics, which tries to understand the neurological basis of how the human reward system affects behavior (Rangel et al., 2008).

### 1.4. Research Questions

In this study we undertake an analysis letting an AI agent learn to manage a renewable resource with two levels of complexity (with or without threshold dynamics). By using this method we can parameterize the LBD process and its related components, and thus we are able to evaluate key learning parameters in relation to optimal control performance. Hence, for studying the limitations and possibilities of LBD for sustainable management of renewable resources, we probe the following questions:

- How does the LBD process respond to different levels of complexity of a renewable resource?
- How do key learning parameters of the LBD process—such as mental model update rate, discounting, hindsight, and experimentation— influence management outcomes?
- Is there a discrepancy between the optimal values of the key learning parameters, depending on the level of complexity of the resource?

The key learning processes are selected in accordance with learning literature on LBD (e.g. Kolb, 1984), within natural resource management literature (e.g. Armitage et al., 2008; Duit and Galaz, 2008; Ostrom, 1990), and the outline of the RL method. Obviously, learning is a much richer phenomenon than we are able to depict here. However, our aim is to analyze the core of the LBD-process, which confines learning in this study. A similar work, using RL, is conspicuously lacking and we aim to provide useful insights to the discourse on the role of learning and decision-making in natural resource management.

### 2. The Model

To make the model less abstract we can envision the setting to be a fishery management problem, where the agent represents a fishing unit (such as a fisherman or an organization with full ownership rights having the ability to exclude other actors, i.e., not a common pool situation), that interacts with a fish stock (where the fish is assessed as a single-species unit). For such situations, ample theory has been developed (Clark, 2010). A list of terminology is provided in Table 1. No a priori understanding of the resource system was given to, nor built into, the agent in order to study the complete learning process (which would theoretically be the case when exploring a new system). Instead, for each fishing event, the agent could set its harvest efforts, observe the harvest, and learn from this experience. The agent was analyzed interacting with either a resource characterized by a logistic growth rate, or a similar resource but with a threshold in its regeneration rate. The two scenarios will further be referred to as the logistic function and the threshold function, to state which dynamics are in focus.

#### 2.1. Resource Dynamics and Agent’s Maximization Problem

First, let us describe the agent’s goal function and the two resource functions. The goal of the agent was to find the effort resulting in the maximum economic yield (MEY) over time. The agent acted as a price-taker in a competitive market, and thus lacked market power.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Terminology and selected parameters used interchangeably throughout the paper.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model term</strong></td>
<td><strong>Fishery term</strong></td>
</tr>
<tr>
<td>Agent</td>
<td>Fisherman/fishing unit*</td>
</tr>
<tr>
<td>Action (a)</td>
<td>Effort/harvest effort</td>
</tr>
<tr>
<td>State (s)</td>
<td>Biomass/stock</td>
</tr>
<tr>
<td>Reward (r)</td>
<td>Net income of harvest</td>
</tr>
<tr>
<td>Time step (t)</td>
<td>Fishing event</td>
</tr>
</tbody>
</table>

* A centrally organized unit of fishermen.
The economic yield was calculated as the value of harvest, minus the cost of harvesting, i.e. the goal function was defined as:

\[ r_t = p q S_t - c q_t \]  

(1)

in which \( r \) was the reward, \( p \) the market price obtainable by the fisher, \( a \) the effort, \( s \) the biomass, and \( c \) the cost of fishing. In our simulations we let \( p = 1.0 \), and \( c = 0.1 \) thus a higher effort generated linearly a higher cost of fishing, and could be interpreted as a simple version of accounting cost.

The resource system was modeled in two ways allowing us to investigate the limits of LBD in relation to resource complexity. For the simpler dynamics we used the Gordon–Schaefer model with linear harvest rate. It is based on a logistic equation and commonly used in resource economics as a general purpose representation of a renewable resource such as fisheries (Clark, 2010). Similarly, a Holling type 2 response function is a general purpose representation of the effect of a vertebrate predator, such as piscivores, on a prey population (Steele and Henderson, 1984). Thus the equations used were:

Logistic function:

\[ s_{t+1} = s_t + g s_t \left( 1 - \frac{s_t}{K} \right) - a s_t - \xi s_t \]  

(2)

Threshold function:

\[ s_{t+1} = s_t + g s_t \left( 1 - \frac{s_t}{K} \right) - q \frac{s_t^2}{h^2 + s_t^2} - a s_t - \xi s_t \]  

(3)

in which \( g \) was the growth rate, \( K \) the carrying capacity, \( s \) the biomass, \( a \) the effort, \( \xi \) the maximum predation rate, and \( h \) set the level where predation was 50% of the maximum predation. The parameter \( \xi \) represented a probability of 0.05 (with uniform random distribution), where 0.05 to 0.95 stock removal was enforced, in order to mimic exogenous events and create a dynamic system.

In Appendix A1 the constants of the above parameters and their corresponding values are stated. The MEY is found by using the following optimal control function max\(0, \frac{1}{2} (1 - s/K) - (K + c)/2s\) - \( q \times s/(h^2 + s^2)\). Note that \( q = 0 \) for the logistic function.

The change in growth is shown in Fig. 1A. In Fig. 1B we show the steady state solution in yield (reward, \( r \)) for any given effort. While the logistic function has one unique solution for the span of effort [0, 1], the threshold function displays two attractors. Should one increase effort above the threshold \( e_1 \), hysteresis is defined as the amount of effort one has to reduce from \( e_1 \) to \( e_2 \) in order to return to the upper and higher yield attractor.

To achieve MEY the state must be kept at \((K + c)/2\). Since the reward is a simple linear function (Eq. (1)), a most rapid approach solution is valid for both resource functions (Clark, 2010). This solution can be readily calculated and is depicted in Fig. 1C. Note that despite the additional threshold term in Eq. (3), the optimal rapid approach of actions are very similar as shown in Fig. 1C. An example of dynamics when following this optimal path of action can be seen in Fig. 1D, where the dips correspond to the enforced stock removal. The optimal control function was approximated by the agent through the learning model described in Section 2.2.

2.2. The Agent

The study of the LBD process on a resource’s complex dynamics required a method that let us study separate details of the important cognitive mechanisms involved. The method needed to entail a learning process, a model of long-term memory and decision-making, and allow for solving future reward problems. While there are multiple ways of modeling dynamic decision-making and learning (see e.g. Busemeyer and Pleskac (2009); Tesfatsion and Judd (2006)), reinforcement learning (RL) approaches, artificial neural networks and the Softmax decision-model have been successfully combined on a diverse set of problems. However, the combination of applying these methods on resource management problems to study aspects of the learning process is, to our knowledge, novel. Below follows an overview of the methods, and the conceptual model in Fig. 2 shows how the agent is constituted of the three methods as well as how they interact with the resource.

RL (Sutton and Barto, 1998) is a selection of problems with a large set of theoretically well-founded learning algorithms, building on ideas of traditional reward based learning. While originating from cognitive psychology it has been brought into the field of machine learning. Lately it has shown to be a promising method for studying human learning in a computerized form in a variety of applications and disciplines (Bogacz et al., 2007; Gureckis and Love, 2009; Montague et al., 2004; Rangel et al., 2008; Wiering and Otterlo, 2012). Temporal difference (TD) learning is a class of algorithms within RL, specialized for prediction i.e., it is able to account for future outcomes by using past experiences within some known or unknown environment. It can be efficiently combined with eligibility traces (Singh and Sutton, 1996) to handling...
The main learning parameters of the agent. maximizing reward. Hence it allows for studying complete exploration versus complete exploitation and the range in between. Note that we use the term exploration level to represent experimentation (as defined in adaptive management; to deviate from perceived optimal action in order to learn about the system) and exploration (as commonly used in e.g. the exploration versus exploitation dilemma). See Table 2, exploration level, for further details.

Together these three parts (TD learning including hindsight, ANN and Softmax policy selection) incorporated the features of learning, planning (prediction and generalization), decision-making and exploration versus exploitation, which specifically allowed for studying important characteristics of these features in a parametric form, particularly the important trade-off between exploration and exploitation; far-sightedness and being myopic; update rate of mental model; and level of hindsight. The parameters are described in detail in Table 2. As shown, these combinations of methods appear highly relevant and rewarding for the theoretical development of learning within the discourse of natural resource management.

2.2.1. Learning Method—Temporal Difference Learning

The key computation in the TD learning algorithm concerns the s.c. TD error between the three values: 1) the immediate reward, 2) the estimated value of the next reward, and 3) the estimated value of the current reward. The estimation is what the agent expects the outcome to be, hence this value is the agent’s mental projection of what to expect in a specific action and state in relation to other states and actions. The estimate of the next reward is weighted by the discount factor, , thus the error not only incorporates immediate rewards but accounts for future earnings. The prediction values (Q values) are retrieved from the agent’s mental model. The equation for the TD error is

\[ \delta_t = r_t + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t). \]  

(4)

In which \( \delta \) was the TD error, \( r \) the reward, \( Q \) the prediction value, \( a \) the action, and \( s \) the state. The error was weighted by the update rate \( (\alpha) \), and then used for adjusting the agent’s mental model \( Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \delta_t \). For further details including the pseudo code for the algorithm used in our simulations, see Appendix A2.

2.2.2. Prediction of Future States

In order to calculate the next prediction \( Q_{t+1}(s_t, a_t + 1) \) in Eq. (4)), an estimate of the future state of the resource is required. Commonly, in TD problems, the actual next state can be provided in the model, but given our conditions, it was not feasible that the agent could know the next state of the stock. Thus, the agent had to make a prediction of the stock after a specific action was taken. In our model we chose to use a simple regression model based on past experiences, commonly used delayed rewards and learn from past experiences. Eligibility traces update the agent’s mental model by back-propagating current outcomes and changing the agent’s mental model of what to expect in relation to recent actions. We call it simply hindsight. TD learning is further described in Section 2.2.1.

Artificial neural networks (ANN) (e.g. Haykin, 2004) have the ability to create computerized models with the main core functions of predictability and generalization. While there are a numerous ways to store the information processed by the RL methods, a radial basis neural network (Poggio and Girosi, 1989) is capable of universal approximation i.e. it has the ability to approximate any given function up to a certain degree, and is used in different memory models (Okamato et al., 2003; Poggio and Bizz, 2004; Tsumori and Ozawa, 2003). Thus, the radial basis neural network can separate knowledge generated from different conditions, while still generalizing into unknown domains. From here on the term mental model will be used for representing the agent’s ANN.

The mental model will create a map of best actions under any given circumstance. However, there are several reasons for not always choosing the best action according to the mental model’s belief. Both the need and desire to experiment, under the assumption that the knowledge held is not perfect, or due to bounded rationality of the agent, would suggest suboptimal actions at times. The Softmax policy selection method (Sutton and Barto, 1998) captures these aspects (Schuster, 2012). It allows for a mechanism to steer the decision-making between the two extremes of 1) complete random choice or 2) the choice leading to

![Fig. 2. Conceptual model of the agent-resource system. The agent architecture consisted of a learning method, a mental model, and a decision-making model. The learning method updated the mental model, and the mental model affected the decision on what action was chosen. The parameters discount factor (γ), level of hindsight (λ), update rate (α), and exploration level (τ), can be set to different values in order to influence the learning process, see Table 2.](image-url)
for prediction purposes, see e.g. (Myers, 1990). This meant that the agent had the ability to understand 1) that the resource system was self-generating and 2) that it affected the resource system by harvesting (Eq. (5)). Letting \( a \) represent the action, \( s \) the state, and \( \mu \) a regression constant, then the prediction of the next state was calculated as
\[
s_{t+1} = \mu_1 s_t + \mu_2 a_t s_t.
\]

### 2.2.3. Mental Model

The mental model's main task was to generalize and store knowledge (the memory of the agent). An important property of the mental model is the feature of generalization versus specific action/state space knowledge. As the agent acquires its first experiences, the agent generalizes the gained knowledge through his whole action/state space to some degree. As more experiences are gathered, the generalization is lowered, and the agent accumulates more specific knowledge in different areas of the action/state space. Hence, a more nuanced mental model of the resource dynamics is formed over time.

The mental model consists of matrix of Gaussian functions in two dimensions, one dimension for state and one for actions. Too few Gaussian functions yield a narrow mind lacking the ability to differentiate between different actions and states, while too many Gaussian functions hinder generalization. The Gaussian functions were evenly positioned, which allowed the agent to have equally speciﬁc knowledge. The Gaussian functions were evenly positioned, which allowed the agent to have equally speciﬁc knowledge.

### 2.3. Computer Experiments

In our simulations each time step corresponded to a specific discrete time frame, for example a week. The model was run 300 time steps and subsequently, if a time step was interpreted as one week then the period of analysis would correspond to approximately 6 years (300/52), and so forth. During one time step the agent 1) made a decision 2) performed an action 3) received a response and 4) updated its mental model of the resource dynamics. See Table 3 for details on these four steps and conceptual model Fig. 2. Notably, 300 time steps is far from infinite, however our test runs showed negligible effects of longer simulations and hence we chose 300 steps to save valuable computing time. Simulations and analysis were performed using MATLAB (2013). Details on the simulations can be found in Appendix A.5.

### 2.4. Performance Index

The evaluation criteria for analyzing the performance of each agent were calculated by a performance index. This performance index was the sum of the agent’s rewards (Eq. (1)), divided by the optimal reward, during T time steps;

\[
PI = \frac{T}{t} \sum_{t=1}^{T} r_t
\]

where \( t \) is the time step, \( r \) is the agent’s reward, and \( r^* \) is the reward given by the analytical solution. Since we enforce some mortality rate, the optimal reward is adjusted to the corresponding drop in reward this causes. Further note that the reward for the threshold function has the same ability to reach a performance of 1.0 despite its slow recuperation rate.

### 3. Results and Discussion

This section addresses the following two areas: 1) a more detailed description of the results and a discussion of the methodological issues of this model, and 2) a more speculative section aiming at contextualizing the findings and their implications for management.

#### 3.1. Learning Parameters and Methodological Issues

##### 3.1.1. Discount Factor

For optimal performance, discount factors of 0.95 and 0.65 were found for the logistic and threshold function respectively, hence future rewards were valued at 95% versus 65% (Fig. 3A). The performance dropped off sharply when deviating from this for lower discount factors but showed little effect for an increase in discount factor for the logistic decision when presented with a loss or gain in relation to a specific decision. However, the level of exploration versus exploitation can be set, thus we can steer the agent’s risk taking behavior from one extreme of little (or no) risk taking, a risk-averse agent, to the other extreme of a risk-seeking agent, where the agent does not discriminate between choices in relation to outcome. Thus, even if we do not model risk preferences directly in the agent’s utility function, it is captured indirectly.

### Table 3

<table>
<thead>
<tr>
<th>Reality</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1 choose action Decide on effort, e.g. how many nets to put out, or fishing fleets to send out, in order to maximize the harvest over time.</td>
<td>The agent decides on an action using the Softmax decision model.</td>
</tr>
<tr>
<td>Step 2 perform action Go fish, e.g. place the nets, or send out the fishing fleets.</td>
<td>The agent applies its chosen action.</td>
</tr>
<tr>
<td>Step 3 receive catch Receive the harvest and calculate the net income.</td>
<td>The resource’s new stock and harvest is calculated (Eqs. (2) and (3)).</td>
</tr>
<tr>
<td>Step 4 learn Learning. Validate the outcome against the expected outcome.</td>
<td>The agent calculates its reward (Eq. (1)).</td>
</tr>
<tr>
<td></td>
<td>The agent updated its mental model using temporal difference learning.</td>
</tr>
</tbody>
</table>
function, while the threshold function experienced a sharp drop for higher values as well. Our results showed that the discount factor was very efficient; with zero discount factor, using the logistic function, the agent only achieved half of the performance while for the threshold function, performance dropped down to just above zero. The optimal values found for the discount factor, 0.95 and 0.7, are in agreement with empirically derived discount factors found in experimental cash flow experiments (Benzion et al., 1989, 276).

3.1.2. Hindsight
The hindsight parameter explicitly accounts for backtracking the actions that have led up to the current situation and reevaluates these according to the current reward. The degree of hindsight (Fig. 3B) (i.e. the number of steps back-traced) did not affect performance considerably for the logistic resource function between 10% in performance, but for the threshold function the result dropped considerably for the threshold function between 10% in performance, but for the threshold function the result dropped sharply. Interestingly, for the threshold function, hindsight improved performance until $\lambda = 0.7$, followed by the same sharp drop. This initial increase in performance fits well with the fact that recovering from a shift into a low productive attractor requires a relatively long time to recover, and thus rewards for choosing a low effort are delayed and need to be back-tracked. However, back-tracking past experiences too far, and thus overvaluing their importance compared to newer experiences, had remarkably negative impacts on the outcome. This result enlightens a dilemma of what timeline to acknowledge when accounting for past experiences.

3.1.3. Update Rate of the Mental Model
The update rate determines the balance between retaining and overwriting knowledge. The balance depends on quickly learning the perceived best course of action (high update rate) against preserving and avoiding overwriting already established knowledge (low update rate). The results (Fig. 3C) showed a blunt peak in performance index at $\alpha = (0.1, 0.5)$ for the logistic function and $\alpha = 0.25$ for the threshold function, although within 10% of maximum performance between $\alpha = (0.1, 0.4)$ for both problems. Hence, the agent had a robust span in its update rate, in which near optimum was achieved. However, for update rates close to zero, the results were considerably different between the functions. For the logistic function the effect showed a slight decrease of 10% in performance, but for the threshold function the result dropped down to a negative outcome. This is related to the exploration level and the shape of the utility function. For further information see discussion in Section 3.1.4.

More effective learning could possibly be achieved by adapting the update rate over time to outcomes deviating from the expected, e.g. uncertainty or surprise, which might further improve the agent’s performance (see e.g. Bowling and Veloso, 2002). In relation to risk behavior, interesting future work could entail inserting additional parameters for updating the agent’s mental model in order to value the outcome of different actions in relation to perceived loss or gain, i.e., more in line with prospect theory.

3.1.4. Exploration Level
The test values for the exploration level (Fig. 3D) had a logarithmic scale from 0.001 to 10. Both functions shared an increase in
performance from low values starting at 0.001 up to an exploration level of 0.05 for the threshold function and 0.6 for the logistic function. However, the exploration level for the threshold function experienced a sharp drop to negative performance while only a slight decrease in performance was observed for the logistic function. The low sensitivity to high exploration values of the logistic function depended on the quadratic shape of the utility function at different actions (Fig. 1B); close to the optima small changes in action did not have a large effect on the reward, and the resource system compensated slightly for over/under harvesting by subsequent reduction/increase in stocks during the next time step. However, the tipping point of the threshold function was close to the MEY and thus repeated unplanned exploration close to MEY but above the tipping point led to entering the low yield attractor.

3.1.5. General Results

Here we show that, although a learning process is costly due to the need for experimentation, the learned behavior can result in 90% (logistic function) or 65% (threshold function) efficiency of the theoretical optimum (Fig. 3). Varying the learning parameters showed a range in performance from negative values (where the cost of harvesting was higher than the net income) to 90% (logistic) and 65% (threshold) of optimal performance, hence the combination of parameter values was very important. The optimal learning parameters we found for the two resource functions indicate that the agent's optimal learning in solving the logistic function was characterized as highly exploitative but less sensitive to changes in the other learning parameters. On the other hand, the agent's behavior when harvesting a resource with a threshold present was characterized by careful exploration in combination with stronger dependence to the other learning parameters. Exploiting a resource with a threshold effect was more difficult to learn, due to being a function of one higher degree and also because its internal dynamics were more complex (the hysteresis effect) as Fig. 1B depicts. Thus, even though the resources' optimal control functions were very similar (Fig. 1C), the threshold function was also harder to learn and to manage using a LBD process.

For the threshold function the importance of the learning parameter values was evident because of the sensitivity to deviations from the optimal values, which showed a rapid drop in performance as one moved away from each fixed optimal parameter value (Fig. 3). The reasons for this disparate behavior were related to the shape of utility function, Fig. 1B, as discussed in Section 3.1.4.

Most reinforcement learning approaches have the property to theoretically converge to an optimal solution over time (Sutton and Barto, 1998). However, a general property of optimal control of complex systems is that while robustness or efficiency is generated for a particular situation, fragility is thereby created for another (Anderies et al., 2007; Csete and Doyle, 2002). Our results confirm this to some degree since the optimal parameterization for a logistic resource function would create severe penalties if the system transformed into a threshold resource function, a situation that could potentially occur in reality if a previously predator-free population became invaded by a predator (corresponding to the predation term in Eq. (3)). Particularly, the optimal experimentation level for the predator-free resource function would drive a predation driven resource to extinction. In a forthcoming paper, we will analyze how changes in population growth rate potentially create trade-offs in learned management behavior.

While there are multiple ways of modeling managerial behavior, our model includes updating mental models, discounting future outcomes, a trade-off in exploration versus exploitation, and hindsight, all relevant in resource management dilemmas. The following discussion touches upon management in relation to these issues.

3.2. Implications for Governance and Management

Our results in Fig. 3 highlight that the learning parameters of the individuals managing the resource stock are crucial in a LBD setting. This poses the question of whether these learning parameters could actually be choices “set” by managers, or if they are constraining personality traits. While we are hesitant to suggest either, we hypothesize that learning traits are possible to influence using different policy instruments. Consider, for example, common strategies affecting management such as education, institutional arrangements e.g. tradable licenses, or social structures. Education, such as gaining understanding of fish ecology, could potentially affect both hindsight (realizing how actions in the past most likely affected current stocks), and discount rate (understanding that high future stocks actually provide the basis for regeneration) (Akpali, 2008; Krasny et al., 2010). Tradable licenses may strongly affect discount rates, as the future value of a license is directly dependent on the potential rent from using it. Horizontal social structures may provide conditions for knowledge sharing (and thereby increasing the pool of experiences from experimentation) while hierarchical organizations with strong leadership may provide central knowledge and make some individuals’ learning traits more pervasive (Berkes, 2009; Crona and Bodin, 2010). Both horizontal and hierarchical social structures could potentially increase the potential for reevaluating ones own mental model of the systems’ dynamics, and for encouraging either exploration or exploitation. Subsequently, whether the learning parameters are considered choices or traits, they are able to be influenced to some degree, although possibly constrained due to other factors such as forced extraction due to lack of options (Ostrom, 1990).

Learning has a cost in that to gain understanding one has to venture into the unknown, and try what one may perceive as suboptimal behavior under the assumption that the knowledge held might be wrong or outdated, and experimentation might lead to better understanding and a long term increase in rewards (Walters and Holling, 1990). In most western and developed countries fishery management is done by scientifically informed government set quotas (Clark, 2007). Science is a learning process in itself relying on combinations of fishery harvest data, independent stock assessments, and biological studies, to test hypotheses of their more or less complex interactions. Just as individual knowledge is continuously updated so is the scientific knowledge base when confronted with actual data (Angelini and Moloney, 2007). In our simulations an update rate span between 0.1 and 0.4 indicates that a substantial degree of overwriting current knowledge, with new knowledge gained from experience, is beneficial. Whether current update rates in fishery science are adapted to the changes actually ongoing in nature, or whether local knowledge is capable of better tracking these changes, remains an open question.

Adaptive management is currently used to infuse the notion of local experimentation (i.e. deviating from the predicted optimal path based on current fisheries models) for learning in fishery governance (Berkes, 2003; Hilborn, 2007). Notably these exploration rates are affecting the total stock and thus the total harvest of the system. In a real world case, these experiments would typically be localized (spatially and temporally), and thus only affecting parts of the resource and corresponding harvests, reducing the overall risk of the experiment (Gelich et al., 2009; Lam and Ostrom, 2010). Subsequently, high exploration rates might be less costly on average because they provide increased benefits in overall system understanding, and thus generate a higher aggregated performance as shown in e.g. Bodin and Norberg (2005). On the other hand, if experimenting on a larger scale, trial-and-error approaches that await evidence of negative environmental impacts before taking action are ill-advised (Biggs et al., 2009).

Given the rather high performance of these naïve agents, when equipped with simple learning algorithms, one question begs to be asked: why does it not work this well in real life? First, the setting we have analyzed is not representative of most open access systems that do show significant problems with e.g. rule compliance or management agreements, in achieving sustainable use (Gutiérrez et al., 2011; Ostrom, 1990). Second, the actor’s ability to control ecological fluxes—unless concerned with a closed system, such as a lake, or a non-migratory specie—are severely limited. Further, the involvement of
many actors changes the reward and learning function as a) the feedback of the expected future rewards (as used in the learning algorithm) is removed since predictability of the future stock is no longer a function of one's own effort only, and b) total effort is determined by entry/exit dynamics that are based on the possibility of positive rent from the fishery. However, local communities may be able to enforce excludability. They are often hierarchically structured to some degree allowing leaders to be central in the information network where they can gather experience and knowledge from others, as well as having social influence on the other fishers in the community (Gutiérrrez et al., 2011). This may provide the conditions for effective learning by more or less a single agent as shown in this paper. The model we used here was designed to investigate the limits of learning optimal control in an isolated learning situation, but the framework lends itself to further studying the role of learning in an open access setting, which is also the goal of our future research.

4. Conclusion

Optimal control management of renewable resources is difficult to obtain because the systems from which they are generated are inherently complex and uncertain. As a result of this unpredictability, a continuous learning process is highly emphasized (Anderies et al., 2007; Levin et al., 2012; Polasky et al., 2011). Although learning is a vast concept, we show that learning, as defined in this context, can be successful. However, efficiency decreases when thresholds are introduced. Further, our results display some critical discrepancies in the process of LBD between key learning parameters, in relation to the degree of complexity of a resource. 1) When there is a possibility of a regime shift, a more greedy policy is beneficial given the condition that exogenous shocks will drive the system into a less productive regime. 2) Hindsight enhances the performance by 20% when thresholds are introduced emphasizing the need to learn from previous experiences when there is a possibility of a regime shift. 3) A mental influence on the simple scenario would propel the stock to extinction when there is a risk of crossing a tipping point. This reveals the fragility of the parameter setting between these two scenarios and suggests finding a “best on average” setting for the parameters to study the performance of LBD in a robust setting, however, we leave this for future research.

In addition, we found a higher sensitivity in the learning parameters when deviating from fixed optimal values in regard to performance between a resource with threshold dynamics, vis-à-vis a resource deficiency. This implies that resources with the possibility of entering a new regime require careful attention to management decisions and entail a substantial body of knowledge to avoid mismanagement. Therefore, this study confirms, in accordance with the results of Biggs et al. (2009), the inherent problems with applying trial-and-error methods in a resource management situation where thresholds are present because of the high costs associated with trespassing a tipping point. On the other hand, if adaptive management would be practiced in systems with similar configurations, and as part of an adaptive management program (Walters, 2007), increased benefits in overall system understanding could be achieved, thus generating a higher aggregated performance of theses systems as shown in Bodin and Norberg (2005).

Acknowledgments

We are grateful to Örjan Ekeberg (KTH Royal Institute of Technology) for help on the methodological design, Antonia Nyström Sandman, Caroline Schill, Andreas Hormegård, Therese Lindahl, Nanda Wijermans, Maja Schlüter, Andrew Merrie, and Juan Carlos Rocha for discussions and constructive comments on earlier versions of the manuscript, and three anonymous reviewers for their valuable suggestions. This research was supported by Mistra through a core grant to the Stockholm Resilience Centre, Stockholm University.

Appendix A. Model Details and Specifications

NB. Abbreviations: reinforcement learning (RL), temporal difference (TD), maximum economic yield (MEY).

Appendix A.1. Parameters for the Resource Functions and Goal Function

Parameters used in the logistic and threshold functions Eqs. (2) and (3), and for the goal function of the agent Eq. (1), are shown in Table A.1.

Appendix A.2. The Learning Algorithm

We implemented a modified version of the on-line Sarsa (λ) algorithm (Rummery and Niranjan, 1994). Modifications were primarily necessary to adjust for continuous states, secondly to implement a

Table A.1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Log</th>
<th>Thr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrying capacity (K)</td>
<td>1.0</td>
<td>1.024</td>
</tr>
<tr>
<td>Growth rate (g)</td>
<td>1.0</td>
<td>1.4784</td>
</tr>
<tr>
<td>Maximum predation rate (q)</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Half saturation constant (b) sets the level where predation is 50% of the maximum predation</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Price for fish (p)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Cost of going fishing (c)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The learning algorithm. The function goFish calls the resource (logistic or threshold function), calcSoftmax calculates the agent’s decision according to the Softmax policy, NeuralNet retrieves the prediction value (Q) for a specific action and state, trainNeuralNet trains and updates the neural network (the agent’s mental model). Parameters: a, action, s, state, e, eligibility trace, δ, TD error, γ, discount factor, and α, update rate.

SET $T$ to number of time steps

SET $t$ to 1

SET $a_t$ to a random number between 0 and 1

SET $s_t$ to a random number between 0 and 1

REPEAT

CALL goFish with $a_t$ RETURN next state $s_{t+1}$

CALL predictNextState with 30 latest states RETURN $p_{s_t+1}$

CALL calcSoftmax with $p_{s_t+1}$ RETURN $a_{t+1}$

Compute reward $r_t$ as $a_t s_t - a_t ^* 0.1$

CALL NeuralNet with $s_t$, $a_t$ RETURN $Q_t$

CALL NeuralNet with $p_{s_t+1}$, $a_{t+1}$ RETURN $Q_{t+1}$

Compute $\delta$ as $r_t + \gamma Q_{t+1} - Q_t$

SET $e(a_t, s_t)$ to 1

SET $i$ to $t$

WHILE $e(a_t, s_t) > 0$ AND $i > 0$

CALL NeuralNet with $s_t$, $a_t$ RETURN $Q_t$

Compute new $Q_t$ as $Q_t + \alpha \delta e(a_t, s_t)$

CALL trainNeuralNet with $a_t$, $s_t$ and $Q_t$

Update $e(a_t, s_t)$ as $\lambda e(a_t, s_t)$

DECREMENT $i$ with one

ENDWHILE

SET $a_t$ to $a_{t+1}$

SET $s_t$ to $s_{t+1}$

INCREMENT $t$ with one

UNTIL $t = T$
more reasonable prediction of future state, and thirdly because the concept of epochs (going back and relearning the same problem) is not relevant for natural resource problems. In Table A.2 the pseudo code for the algorithm is given to provide a detailed specification of the modified algorithm.

Appendix A.3. The Agent's Mental Model and Decision-making (Softmax)

The agent’s mental model (see Fig. A.1) is a map of how he values every action for a specific state. In both cases he learns that a low harvest is more beneficial when states are below 0.55. However there is a difference in how close the agent is to MEY in the logistic versus the threshold function. For the logistic function the agent manages to find the MEY but for lower states the agent does not adapt as efficiently as the agent learning the threshold function. The agent learning the threshold function displays a more careful behavior as its optimal strategy and is less explorative. This careful behavior corresponds to the results of Lindahl et al. (2012) where lab experiments with human subjects indicate a more careful behavior of subjects exposed threshold problems compared to threshold-free problems. Further, the agent commonly overexploits in the logistic function, although only slightly at MEY. However for the threshold function the agent overexploits for states below 0.4 and underexploits for higher states.

Fig. A.2 illustrates the effect of choosing lower or higher values of different exploration levels using the decision-making model, Softmax. Lower values increased the probability of the best perceived action being chosen, while higher values generate a more equal distribution of all actions being chosen. The Softmax algorithm used in our experiments can be found in pseudo code in Table A.3.

Appendix A.4. Sensitivity Analysis Between Number of Gauss Functions and Parameters

Fig. A.3 illustrates the sensitivity of the parameter values in relation to the number of Gaussian functions representing the mental model of the agent. Fig. A.4 visualizes the reduced ability to approximate the discontinuity, and thus the complete optimal control curve, of the optimal control function when the mental model is reduced from 81 Gaussian functions to 36.

Appendix A.5. Simulations

First, we simulated 3000 agents with uniform random combinations of the parameters discount factor, update rate, hindsight, and exploration level to find the parameters of the best agent in the logistic and threshold

Table A.3

The Softmax algorithm. NeurNet calls the mental model of the agent and returns Q. The parameter a is the action, s the state.

CALL NeurNet((a, s) for all a RETURN Q
Calculate Q as Q = max(Q)
Calculate Q as Q^+
SET probability P to Q = Q / sum(Q)
SET C to cumulative probability of P
SET η to a random number between 0 and 1
SET a to action corresponding to C matching η
Fig. A.3. Top panel shows the values for the logistic function, and the bottom panel shows the values for the threshold function. The figures show the sensitivity of parameter values in relation to the number of Gaussian functions used for the mental model of the agent. The solid line depicts the optimal parameter value used to the smoothed function over the data points. The dash–dot line shows the actual optimal parameter value, and the dashed line the optimal parameter value found and used when computing the agent with 81 Gaussian functions as used in the main analysis. The data represents an average of 100 runs.

Fig. A.4. Each mental model depicted learns by receiving random states and the corresponding best actions and rewards, with an update rate of 0.2 and during 300 time steps. Thus this agent learns to map combinations of actions and states to correct values. Blue means less outcome is expected while dark red means a higher outcome is expected. Top panel show the results for 36 Gaussian functions, and the lower panel shows the result of 81 Gaussian functions. The logistic function is shown on left hand side and threshold function on the right hand side.
functions respectively. Every combination was repeated 50 times to minimize the effect of the stochasticity in Softmax. ... An Introduction. NetLibrary, Incorporated.

References


