Exploration in POMDPs

Article *in* OGAI Journal (Oesterreichische Gesellschaft fuer Artificial Intelligence) · March 2008

1 author:

Christos Dimitrakakis
University of Oslo
103 PUBLICATIONS **1,294** CITATIONS

UNIVERSITEIT
VAN
AMSTERDAM

Submitted to: Österreichische Gesellschaft für Artificial Intelligence

IAS technical report IAS-UVA-08-01

Exploration in POMDPs

Christos Dimitrakakis
Intelligent Systems Laboratory Amsterdam,
University of Amsterdam
The Netherlands

In recent work, Bayesian methods for exploration in Markov decision processes (MDPs) and for solving known partially-observable Markov decision processes (POMDPs) have been proposed. In this paper we review the similarities and differences between those two domains and propose methods to deal with them simultaneously. This enables us to attack the Bayes-optimal reinforcement learning problem in POMDPs.
Keywords: POMDPs, exploration, bayesian, reinforcement learning, belief

IAS
intelligent autonomous systems

Contents

Intelligent Autonomous Systems
Informatics Institute, Faculty of Science
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
Tel (fax): +31 20 525 7461 (7490)
http://www.science.uva.nl/research/ias/

Corresponding author:
C. Dimitrakakis
tel: +31 20 525 7517
dimitrak@science.uva.nl
http://www.science.uva.nl/~dimitrak/

1 Introduction

Let us consider the problem of an agent acting in a discrete-time dynamic environment. The dynamics of the environment are such that the transition from the current state s_t to the next, s_{t+1} depends only on the current state and the action a_t the agent is taking at the time. In addition, there exists a reward signal $r_t \in \Re$, and the agent wishes to maximise the expected utility $\mathbf{E}U_t(T)$, with $U_t(T) = \sum_{k=1}^{T-t} g(t+k)r_{t+k}$, where T is the horizon after which we are no longer interested in rewards and $g \geq 0$ is a discounting factor which allows us to express the relative value of rewards in the future. If the agents can observe the state, then the process of the agent's interaction with the environment can be formally defined as follows.

Definition 1 [Markov decision process] *A Markov decision process μ (MDP) is defined as the tuple $\mu = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ comprised of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition distribution $\mathcal{T}$ conditioning the next state on the current state and action, $\mu(s'|s,a) = \mathbf{P}(s_{t+1} = s'|s_t = s, a_t = a, \mu)$ and a reward distribution $\mathcal{R}$ conditioned on states and actions $\mu(r|s,a) = p(r_{t+1} = r|s_t = s, a_t = a)$, with $a \in \mathcal{A}$, $s, s' \in \mathcal{S}$, $r \in \Re$.*

Note that in this definition, and in the rest of the text, we are using the notational convention $z(x|y) \equiv \mathbf{P}(x|y,z)$. The set of all MDPs shall be denoted by $\mathcal{M}$. In the simplest setting, which is usually referred to as dynamic programming [10, 3, 14], we assume that the model μ of the Markov decision process is known and attempt to discover a policy π^* which is optimal for that MDP. A policy π is defined as a distribution over actions conditioned on the state, i.e. $\pi_t(a|s) = \mathbf{P}(a_t = a|s_t = s, \pi)$. The optimal policy given μ, T, g maximises the state value function

$$V^{\pi}_{t,T}(s) \equiv \mathbf{E}(U_t(T)|\pi, s_t = s, \mu) \equiv \sum_{k=1}^{T-t} g(t+k)\mathbf{E}(r_{t+k}|\pi, s_t = s) \tag{1}$$

for every state s in the MDP, i.e. $V^*_{t,T}(s) \equiv V^{\pi^*}_{t,T}(s) \geq V^{\pi}_{t,T}(s)$ for any π, s. The form of the value function determines our objective. When T is finite we are only interested in what occurs in the environment until time T and the problem falls in the category of finite-horizon problems. For the infinite horizon case it is useful to set $g(t+k) = \gamma^{t+k}$, which for $\gamma \in [0,1)$ leads to $\lim_{T\to\infty} V^{\pi}_{t,T} = V^{\pi}$.

In the reinforcement learning framework, μ is not known and we must estimate the optimal policy π^* by interacting with the environment. This framework is described in detail in [3, 14], which provide algorithms for *approximate* dynamic programming. These converge under certain conditions to the same solution as dynamic programming.

However, such approaches mostly deal with the problem from an aspect of optimisation and stochastic approximation theory, while the uncertainty inherent in the problem (we do not know the "true state" of the world), is not directly considered. Typical convergence proofs for such algorithms contain sufficient conditions for approximating an optimal policy that require the agent to act in an exploratory manner in order to reduce uncertainty. Furthermore, although there exist both asymptotic and sample-complexity results, the question of how to behave optimally such as to maximise (1) while learning, had not been addressed in this line of work.

In fact, the question of acting optimally under uncertainty appears even in the simplest possible reinforcement learning setting, the multi-armed bandit problem. There, there is only one state and the agent can take one of a finite set of actions, each of each has an unknown, but usually fixed, mean reward. Depending on the discount parameter γ and the uncertainty about the values of different actions, the agent will choose between a profitable arm or an uncertain, but potentially more profitable one. A set of provably good methods for optimal exploration in this setting is given by [1], with the only restriction being boundedness of the expected rewards. However, in the Bayesian subjectivist setting for sequential decision making [4], there has also

been substantial work towards optimal Bayesian methods for bandit problems, the most well-known of which is the Gittins index [6].

The attractiveness of the Bayesian approach is that, given an initial prior over all possible MDPs, we can create another model, whose state is composed of two sub-states: the state of our belief (a probability distribution over $\mathcal{M}$) and the system state of the original MDP. This belief-augmented MDP can then be solved using standard dynamic programming methods in order to obtain the optimal action under uncertainty. We shall look at the estimation procedure that is involved in creating this MDP, and discuss the cases when this is intractable. In the cases where it is tractable, we shall examine methods for performing optimal action selection by solving the resulting augmented MDP. Because this can be an extremely large MDP, standard solution methods might not be applicable and we will have to resort to approximations.

The rest of the paper is organised as follows. First we shall present belief-augmented MDPs and then extend this formalism to partially observable MDPs. Secondly, we shall present how a belief state can be maintained in either case. Finally, we shall outline current methods for optimal action selection in these settings and propose directions for future research.

1.1 Exploration in MDPs

When we are uncertain about which MDP we are acting in, we may maintain a belief over possible MDPs. If we augment the MDP's state with a belief, we can then solve the exploration easily via standard dynamic programming algorithms such as backwards induction or value iteration. We shall call such models Belief MDPs[1] (BMDPs), analogously to the BAMDPs (Bayes-Adaptive MDPs) of [5]. This is done by not only considering densities conditioned on the state-action pairs (s_t, a_t), i.e.

$$p(r_{t+1}, s_{t+1} | s_t, a_t)$$

but taking into account the belief $\xi_t \in \mathcal{B}$, a probability space over possible MDPs, i.e. augmenting the state space from $\mathcal{S}$ to $\mathcal{S} \times \mathcal{B}$ and considering the following conditional density:

$$p(r_{t+1}, s_{t+1}, \xi_{t+1} | s_t, a_t, \xi_t).$$

More formally, we may give the following definition:

Definition 2 [Belief MDP] *A* Belief MDP ν *(BMPD) is an MDP* $\nu = (\Omega, \mathcal{A}, \mathcal{T}', \mathcal{R}')$ *where* $\Omega = \mathcal{S} \times \mathcal{B}$*, where* $\mathcal{B}$ *is the set of probability measures on* $\mathcal{M}$*, and* $\mathcal{T}', \mathcal{R}'$ *are the transition and reward distributions conditioned jointly on the MDP state* s_t*, the belief state* ξ_t*, and the action* a_t*, such that the following factorisations are satisfied for all* $\mu \in \mathcal{M}$*,* $\xi_t \in \mathcal{B}$*.*

$$p(s_{t+1} | s_t, s_{t-1}, \ldots, s_1, a_t, \mu) = \mu(s_{t+1} | s_t, a_t) \tag{2}$$

$$p(s_{t+1}, \xi_{t+1} | a_t, s_t, \xi_t) = \int_{\mathcal{M}} p(\xi_{t+1} | s_{t+1}, a_t, s_t, \mu, \xi_t) \mu(s_{t+1} | a_t, s_t) \xi_t(\mu) d\mu \tag{3}$$

We shall use $\mathcal{M}_B$ to denote the set of BMDPs. It should be obvious from (3) that s_t, ξ_t jointly form a Markov state in this setting.

The form of the probability distribution over MDPs, $\xi_t(\mu)$, need not be particularly complex. In fact, if we consider discrete states and action spaces, then the belief over transition distributions can be represented with a simple Dirichlet prior for each state-action pair as long as we consider the state-action-state transition distributions to be independent. This is a probability

[1] It is also possible to consider different forms of beliefs than standard Bayesian ones.

distribution over possible multinomial distributions, to which it is conjugate. This is fully characterised by transition counts.[2] More specifically, suppose we have k discrete events, drawn from an unknown multinomial distribution $q \equiv (q_1, \ldots, q_k)$. If $q \equiv (q_1, \ldots, q_k) \sim \mathrm{Dir}(\phi_1, \ldots, \phi_k)$, then the p.d.f. is $\phi(q) \propto \prod_{i=1}^{k} q_i^{\phi_i - 1}$, where ϕ_i is the number of times i has been observed. We shall omit details for the fully-observable case and proceed directly to exploration in partially-observable MDPs.

2 Exploration in POMDPs

A useful extension of the MDP model can be obtained by not allowing the agent to directly observe the state of the environment, but an observation variable o_t that is conditioned on the state. This more realistic assumption is formally defined as follows:

Definition 3 [Partially observable Markov decision process] *A partially observable Markov decision process μ (POMDP) is defined as the tuple $\mu = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R})$ comprised of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition-observation distribution $\mathcal{T}$ conditioned the current state and action $\mu(s_{t+1} = s', o_{t+1} = o | s_t = s, a_t = a)$ and a reward distribution $\mathcal{R}$, conditioned on the state and action $\mu(r_{t+1} = r | s_t = s, a_t = a)$, with $a \in \mathcal{A}$, $s, s' \in \mathcal{S}$, $o \in \mathcal{O}$, $r \in \Re$.*

We shall denote the set of POMDPs as $\mathcal{M}_P$. For POMDPs, it is often assumed that one of the two following factorisations holds:

$$\mu(s_{t+1}, o_{t+1} | s_t, a_t) = \mu(s_{t+1} | s_t, a_t) \mu(o_{t+1} | s_{t+1}) \tag{4}$$

$$\mu(s_{t+1}, o_{t+1} | s_t, a_t) = \mu(s_{t+1} | s_t, a_t) \mu(o_{t+1} | s_t, a_t). \tag{5}$$

The assumption that the observations are only dependent on a single state or a single state-action pair is a natural decomposition for a lot of practical problems.

POMDPs are similar to BMDPs. In fact, BMDPs are equivalent to a special case of a POMDP in which the state is split into two parts: One fully observable dynamic part and one unobservable, but stationary part, which models the unknown MDP. Typically, however, in POMDP applications the unobserved part of a state is dynamic. The problem of acting optimally in POMDPs has two aspects. The first is state estimation, and the second is acting optimally given the estimated state.

As far as the first part is concerned, given an initial state probability distribution, updating the belief amounts to simply maintaining a multinomial distribution over the states. However, the initial state distribution might not be known. In that case, we may assume an initial prior density over the multinomial state distribution. It is easy to see that this is simply a special case of an unknown state transition distribution, where we insert a special initial state which is only visited once. We shall, however, be concerned with the more general case of full exploration in POMDPs, where all state transition distributions are unknown.

2.1 Belief POMDPs

It is possible to create an augmented MDP for POMDP models, by endowing them with an additional belief state, in the same manner as MDPs. However now the belief state will be a joint probability distribution over $\mathcal{M}_P$ and $\mathcal{S}$. Nevertheless, each (a_t, o_{t+1}) pair that is observed leads to a unique subsequent belief state. More formally, a belief-augmented POMDP is defined as follows:

[2] In this sense, it is analogous to the beta distribution, which is the conjugate family to Bernoulli distributions.

Definition 4 [Belief POMDP] *A* Belief POMDP ν *(BPOMPD) is an MDP* $\nu = (\Omega, \mathcal{A}, \mathcal{O}, \mathcal{T}', \mathcal{R}')$ *where* $\Omega = \mathcal{G} \times \mathcal{B}$*, where* $\mathcal{G}$ *is the set of probability measures on* $\mathcal{S}$*,* $\mathcal{B}$ *is the set of probability measures on* $\mathcal{M}_P$*,* $\mathcal{T}'$ $\mathcal{R}'$ *are the belief state transition and reward distributions conditioned on the belief state* ξ_t *and the action* a_t *such that the following factorizations are satisfied for all* $\mu \in \mathcal{M}_P$*,* $\xi_t \in \mathcal{B}$

$$p(s_{t+1}|s_t, a_t, s_{t-1}, \ldots, \mu) = \mu(s_{t+1}|s_t, a_t) \tag{6}$$

$$p(o_t|s_t, a_t, o_{t-1}, \ldots, \mu) = \mu(o_t|s_t, a_t) \tag{7}$$

$$p(\xi_{t+1}|o_{t+1}, a_t, \xi_t) = \int_{\mathcal{M}_P} p(\xi_{t+1}|\mu, o_{t+1}, a_t, \xi_t)\xi_{t+1}(\mu|o_{t+1}, a_t, \xi_t)\, d\mu \tag{8}$$

We shall denote the set of BPOMDPs with $\mathcal{M}_{BP}$. Again, (8) simply assures that the transitions in the belief-POMDP are well-defined. The Markov state $\xi_t(\mu, s_t)$ now jointly specifies a distribution over POMDPs and states.[3] As in the MDP case, in order to be able to evaluate policies and select actions optimally, we need to first construct the BPOMDP. This requires calculating the transitions from the current belief state to subsequent ones according to our possible future observations, as well as the probability of those observations. The next section goes into this in more detail.

2.2 The belief state

In order to simplify the exposition, in the following we shall assume firstly that each POMDP has the same number of states. Then $\xi(s_t = s|\mu)$ describes the probability that we are in state s at time t given some belief ξ and assuming we are in the POMDP μ. Similarly, $\xi(s_t = s, \mu)$ is the joint probability given our belief. This joint distribution can be used as a state in an expanded MDP, which can be solved via backward induction, as will be seen later. In order to do this, we must start with an initial belief ξ_0 and calculate all possible subsequent beliefs. The belief at time $t + 1$ depends only on the belief time t and the current set of observations r_{t+1}, o_{t+1}, a_t. Thus, the transition probability from ξ_t to ξ_{t+1} is just the probability of the observations according to our current belief, $\xi_t(r_{t+1}, o_{t+1}|a_t)$. This can be calculated by first noting that given the model and the state, the probability of the observations no longer depends on the belief, i.e.

$$\xi_t(r_{t+1}, o_{t+1}, |s_t, a_t, \mu) = \mu(r_{t+1}, o_{t+1}|a_t, s_t) = \mu(r_{t+1}|a_t, s_t)\mu(o_{t+1}|a_t, s_t). \tag{9}$$

The probability of any particular observation can be obtained by integrating over all the possible models and states

$$\xi_t(r_{t+1}, o_{t+1}|a_t) = \int_{\mathcal{M}_P} \int_{\mathcal{S}} \mu(r_{t+1}, o_{t+1}|a_t, s_t)\xi(\mu, s_t). \tag{10}$$

Given that a particular observation is made from a specific belief state, we now need to calculate what belief state it would lead to. For this we need to compute the posterior belief over POMDPs and states. The belief over POMDPs is given by

$$\xi_{t+1}(\mu) \equiv \xi_t(\mu|r_{t+1}, o_{t+1}, a_t,) \tag{11}$$

$$= \frac{\xi_t(r_{t+1}, o_{t+1}, a_t|\mu)\xi_t(\mu)}{\xi_t(r_{t+1}, o_{t+1}, a_t)} \tag{12}$$

$$= \frac{\xi_t(\mu)}{Z} \iint_{\mathcal{S}} \mu(r_{t+1}, o_{t+1}, a_t|s_{t+1}, s_t)\xi_t(s_{t+1}, s_t|\mu)\, ds_{t+1} ds_t, \tag{13}$$

[3] The formalism is very similar to that described in [11], with the exception that we do not include the actual POMDP state in the model.

where $Z = \xi_t(r_{t+1}, o_{t+1}, a_t)$ is a normalising constant. Note that $\xi_t(s_{t+1}, s_t|\mu) = \mu(s_{t+1}|s_t)\xi_t(s_t|\mu)$, where $\xi_t(s_t|\mu)$ is our belief about the state in the POMDP μ. This can be updated using the following two steps. Firstly, the filtering step

$$\xi_{t+1}(s_t|\mu) \equiv \xi_t(s_t|r_{t+1}, o_{t+1}, a_t, \mu) \tag{14}$$

$$= \frac{\mu(r_{t+1}, o_{t+1}|s_t, a_t)\xi_t(s_t|\mu)}{\xi_t(r_{t+1}, o_{t+1}|a_t, \mu)}, \tag{15}$$

where we adjust our belief about the previous state of the MDP based on. Then we must perform a prediction step

$$\xi_{t+1}(s_{t+1}|\mu) = \int_{\mathcal{S}} p(s_{t+1}|s_t = s, \mu)\xi_{t+1}(s_t = s|\mu)\, ds, \tag{16}$$

where we calculate the probability over the current states given our new belief concerning the previous states. These predictions can be used to further calculate a new possible belief, since our current belief corresponds to a distribution over possible MDPs. We use the probability distribution over MDPs, and for each possible MDP we determine how our beliefs would change as we acquire new observations. The main difficulty is maintaining the joint distribution over states and POMDPs. This will be further discussed in the final section.

2.3 Action selection

The second difficulty in the exploration task is action selection. For this, we need to select the action maximising

$$V^{\pi^*}_{t,T}(a_t, \xi_t) = \int_{\mathcal{M}} \int_{\mathcal{S}} \xi_t(\mu, s_t) \max_{\pi} \mathbf{E}(U_t|\pi, s_t, a_t, \xi_t, \mu) ds_t d\mu.$$

This is far from a trivial operation. However, in finite-horizon problems we can perform a backwards induction procedure where we start from the optimal action at the last stage and calculate $V^{\pi^*}_{T,T}(a|\xi_T)$ for all possible belief states ξ_T at that stage and we subsequently calculate $V^{\pi^*}_{T-1,T}(a|\xi_T), V^{\pi^*}_{T-2,T}(a|\xi_{T-2}), \ldots, V^{\pi^*}_{t,T}(a|\xi_t)$. Note that at the current time t there is only a single belief ξ_t. Further, at each stage n we can write

$$V^{\pi^*}_{n,T}(a_n|\xi_n) = \int_{\mathcal{O}} \int_{-\infty}^{\infty} \xi_n\left(r_{n+1}, o_{n+1}|a_n\right) \left[r + V^{\pi^*}_{n+1,T}\left(a^*_{n+1}|r_{n+1}, o_{n+1}, a_n, \xi_n\right)\right] dr_{n+1}\, do_{n+1},$$

where $V^{\pi^*}_{n+1,T}(a^*_{n+1}|r_{n+1}, o_{n+1}, a_n, \xi_n) \equiv V^{\pi^*}_{n+1,T}(a^*_{n+1}|\xi_{n+1})$ for one of the possible next-step beliefs. The implication is that we first must calculate all possible belief states starting from the current state, for all stages until T. The complexity of this operation is high, since if at each stage there are n possible observations, then the number of possible beliefs is of order n^T; if the reward, state, action, or observation spaces are continuous, we are presented with a potentially insurmountable problem.

3 Current and future research

The BMDP as presented herein is essentially identical to the Bayes-adaptive MDP formalism use in [5], but the general idea has been around since [2]. Current research focuses on practical methods for decision making in these cases.

The work by [8] is one of the first modern works on POMDPs where uncertainty about the model is explicitly taken into account. In this setting, it is possible to directly query the

true value of the state parameter of the POMDP with a special query action. It is possible to generalise the 'query action' idea by [8] to the case where a query action is just a temporally extended [15] 'exploration' action. As was suggested for example in [**?**], it is possible to place upper bounds on the value of exploration by lower-bounding the regret incurred while taking exploratory movements. Methods for bounding the value of different actions will be particularly applicable to continuous state, action or observation spaces.

An extension of the Bayes-adaptive MDP framework to the POMDP case was also discussed in [11]. The additional problem in POMDPs is that tracking the joint belief over $\mathcal{S} \times \mathcal{M}_P$ is difficult. In most cases the effort concentrates on reducing the amount of states that must be tracked. For standard POMDP problems, fixed point approximations work well [13]. However, exploration problems create the need for a continuous refinement of the discretisation. Monte-carlo methods [7, 9] offer a potential solution. For uncertain POMDP problems, methods such as those used in [11] rely on procedures for compacting the belief space.

Bayesian approaches have not been limited to exploration problems. For example, [16] solve known POMDPs, where a version of the EM algorithm is used to find policies. One of the main ideas therein was considering an infinite horizon MDP as a infinite mixture of finite horizon MDPs. This idea is also used by [7], which extends the procedure to a full Bayesian approach and continuous spaces. Furthermore, they introduce an efficient transdimensional Markov chain Monte-Carlo procedure, which is illustrated to perform at a level comparable to that of the full Bayesian approach. Such methods might be applicable to to the full exploration problem as well.

While solutions methods for known POMPDs could be applied to BMDPs, or BPOMDPs, most currently used methods perform off-line calculations that are geared towards solving a tracking problem and thus they are not directly suitable for exploration. However, recently, progress has been made towards applicable online methods. In particular, [12] offers a theoretical analysis of heuristic online POMDP planning algorithms which shows their near-optimality. It should be possible to apply such methods to uncertain POMDP problems as well.

Multiple agent problems, especially decentralised POMDPs, are an interesting scenario. The question then becomes one of determining how to best explore the POMDP given the additional exploration value that multiple agents add. In particular, it would be of interest to see how the complexity of exploration reduces as the number of agents increases.

In summary, there are three main possibilities for future research. The first is to consider classes of problems with special structure; this might either allow a more efficient solution, or it might lead to an interesting secondary problem. This in turn could result in new theoretical analyses. The second is to consider approximate methods for POMDPs using either high-probability bounds on the value function for parts of the tree in order to perform pruning, Monte Carlo sampling to selectively expand parts of the tree, projection of the belief to a more compact representation, or other heuristics for simplifying computations. Methods for solving POMDPs could be applied to BMDPs. Finally, examining different belief representations than the full Bayesian ones might lead to a more manageable problem.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002. A preliminary version has appeared in *Proc. of the 15th International Conference on Machine Learning*.

[2] Richard Ernest Bellman. *Dynamic Programming*. Princeton University Press, 1957. Republished by Dover in 2004.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[4] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970. Republished in 2004.

[5] Michael O'Gordon Duff. *Optimal Learning Computational Procedures for Bayes-adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

[6] C. J. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, New Jersey, US, 1989.

[7] Matthew Hoffman, Arnaud Doucet, Nando De Freitas, and Ajay Jasra. Bayesian policy learning with trans-dimensional mcmc. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[8] R. Jaulmes, J. Pineau, and D. Precup. Active Learning in Partially Observable Markov Decision Processes. *European Conference on Machine Learning*, 2005.

[9] Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. Reinforcement learning in continuous action spaces through sequential monte carlo methods. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[10] Marting L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New Jersey, US, 1994,2005.

[11] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive POMDPs. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[12] Stephane Ross, Joelle Pineau, and Brahim Chaib-draa. Theoretical analysis of heuristic search methods for online POMDPs. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

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

[14] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[15] Richard S. Sutton, Doina Precup, and Satinder P. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

[16] Marc Toussaint, Stefan Harmelign, and Amos Storkey. Probabilistic inference for solving (PO)MDPs. Research report, University of Endinburgh, School of Informatics, 2006.

Acknowledgements

Frans Oliehoek and Ronald Ortner for useful discussions and proof reading

IAS reports

This report is in the series of IAS technical reports. The series editor is Bas Terwijn (bterwijn@science.uva.nl). Within this series the following titles appeared:

F.A. Oliehoek and N. Vlassis and M.T.J. Spaan, *Properties of the QBG-value function* Technical Report IAS-UVA-07-04, Informatics Institute, University of Amsterdam, The Netherlands, August 2007.

G. Pavlin and P. de Oude and M.G. Maris and J.R.J. Nunnink and T. Hood *A Distributed Approach to Information Fusion Systems Based on Causal Probabilistic Models.* Technical Report IAS-UVA-07-03, Informatics Institute, University of Amsterdam, The Netherlands, July 2007.

P.J. Withagen and F.C.A. Groen and K. Schutte *Shadow detection using a physical basis.* Technical Report IAS-UVA-07-02, Informatics Institute, University of Amsterdam, The Netherlands, Februari 2007.

All IAS technical reports are available for download at the ISLA website, http://www.science.uva.nl/research/isla/MetisReports.php.