
Exploration control in Reinforcement Learning using Optimistic Model Selection

Jeremy L. Wyatt

JLW@CS.BHAM.AC.UK

School of Computer Science, University of Birmingham, Birmingham, B15 2TT, United Kingdom

Abstract

This paper presents two new algorithms for exploration control in reinforcement learning (RL) based on optimistic model selection (OMS) from a density over possible models. We relate this to the Bayesian formulation of the problem, and to a number of recent methods. An empirical study shows that when optimised OMS usually outperforms current model-based methods on three tasks. We discuss the problem of parameter selection.

1. Introduction

The problem of how to act while learning is a class of optimal control problems with a long history (Gittins, 1989; Bellman, 1961). In RL it has taken the form of problems of (i) how to act so as to maximise performance during the learning agent’s lifetime (Martin, 1967; Kaelbling, 1990); and (ii) how to act to identify as good a policy as possible within the learning period (Kearns & Singh, 1991; Fiechter, 1997; Dearden et al., 1999). These problems, while related, are not the same (Wyatt, 1997).

In Markov decision processes (MDPs) the optimal Bayesian solution to problem (i) is well known, but intractable (Martin, 1967; Bellman, 1961). Many approximations have been proposed. In this paper a new heuristic method is presented which brings together ideas from several recent approaches (Kaelbling, 1990; Wiering & Schmidhuber, 1998; Kearns & Singh, 1991). The algorithms derived, when optimised, outperform existing methods on both problems (i) and (ii). The domain is a finite state MDP with an unknown transition function and a known reward function. All the methods considered here are model-based. The paper is structured as follows. In Section 2 we describe the optimal solution within a Bayesian framework, as originally outlined by (Martin, 1967; Bellman, 1961), and recent heuristic approaches. In Section 3 we describe two new algorithms which are ad-

vances on Wiering’s (1998) model based interval estimation (MBIE) method. The main difference is that we use a Dirichlet density, and are thus able to use OMS to guide exploration from the outset. This also gives us a clearer relationship to the Bayesian view of the exploration-exploitation trade-off. In Sections 3 and 5 we present the results of an empirical study, and show that the new algorithms usually outperform both Wiering’s and Meuleau’s (1999) exploration methods on standard tasks when all algorithms are optimised.

2. Previous Work

There are four components of an exploration method (Wyatt, 1997): the measure of *local* exploratory value (e.g. based on reward, counter, error, recency, or variance); whether this measure is converted into a *distal* measure of exploratory value using a Bellman equation; whether the method for inferring the exploration value function is *model-based* or *model-free*; and what form the *decision rule* based on the exploration value function takes (e.g. ϵ -greedy, Boltzmann, deterministic). All the methods considered here are model-based, distal and deterministic. We first outline the Bayesian formulation of problem (i), for the case of an unknown MDP. All other exploration measures for problem (i) can essentially be considered an approximation to this.

The Bayesian approach is based on there being a space \mathcal{P} of possible transition functions (or models) P for the MDP, and a well-defined prior probability density over that space. The probability density over the space of possible finite state MDPs for a known state space \mathcal{S} is constructed as follows. First let us think about the density over the possible one-step transition functions from a single state action pair. If state $i \in \mathcal{S}$ has N possible succeeding states when action a is taken, then the transition function from that state action pair is a multinomial distribution over the outcomes:

$$\vec{p}_i^a = \{p_{i1}^a, p_{i2}^a \cdots p_{iN}^a\} \quad (1)$$

The possible transition functions from i, a are the possible \vec{p}_i^a . We want a probability density over this space

which is closed under sampling from any such multinomial. The Dirichlet density has this property:

$$f(\vec{p}_i^a | \vec{m}_i^a) = \frac{\Gamma(\sum_{j=1}^N m_{ij}^a)}{\prod_{j=1}^N \Gamma(m_{ij}^a)} \prod_{j=1}^N (p_{ij}^a)^{m_{ij}^a - 1} \quad (2)$$

the density is parameterised by the $m_{ij}^a > 0$ for all j . The parameter vector is updated as follows, if a single observation of a transition $i \xrightarrow{a} j$ is made, then the new density is also Dirichlet with $m_{ij}^{a''} = m_{ij}^a + 1$. The density for the multi-state case follows directly from this since the densities over the one step transition functions for all state action pairs are independent. The density $f(P|M)$ for a possible transition function $P \in \mathcal{P}$ for the MDP is therefore simply the product of the $f(\vec{p}_i^a | \vec{m}_i^a)$ over all i . This density is parameterised by the matrix $M = [m_{ij}^a]$, where $M \in \mathcal{M}$. In a Bayesian framework we choose a prior matrix M' , which specifies our prior density over the space of possible models. The additional information from a sequence of observations is captured in a count matrix F . The posterior density given these observations is therefore simply parameterised by $M'' = M' + F$. For convenience the transformation on M due to a single observed transition $i \xrightarrow{a} j$ is denoted $T_{ij}^a(M)$. The value function in an MDP with unknown transition probabilities is thus itself a random variable, \tilde{V}_i . Given the usual squared error loss function the Bayesian estimator of expected return under the optimal policy is the expectation of \tilde{V}_i :

$$V_i(M) = \mathbb{E}[\tilde{V}_i | M] = \int_{\mathcal{P}} V_i(P) f(P|M) dP \quad (3)$$

where $V_i(P)$ is the value of i given the transition function P . The central result of both Bellman and Martin was that when this integral is evaluated we transform our problem into one of solving an MDP with known transition probabilities, defined on the information space $\mathcal{M} \times \mathcal{S}$:

$$V_i(M) = \max_a \left\{ \sum_j \bar{p}_{ij}^a(M) (r_{ij}^a + \gamma V_j(T_{ij}^a(M))) \right\} \quad (4)$$

where $\bar{p}_{ij}^a(M)$ is the marginal expectation of the Dirichlet, $0 \leq \gamma < 1$ is the discount rate, and r_{ij}^a is the reward associated with the transition $i \xrightarrow{a} j$. This shows how the Bayesian estimate of value elegantly incorporates the value of future information. The optimal solution to the well-known exploration-exploitation trade-off (problem (i) above) is thus to act greedily with respect to the Bayes Q-values. Because the solution involves dynamic programming over a tree of information states the problem is intractable. A simple approximation to this is the certainty equivalent (CE) estimate constructed by replacing $T_{ij}^a(M)$

with M in (4). We could also approximate the value of the integral by random sampling (Dearden et al., 1999; Strens, 2000).

Approximate approaches to the exploration-exploitation trade-off typically circumvent this problem by some instantiation of the heuristic “be optimistic in the face of uncertainty” (Moore & Atkeson, 1993; Kaelbling, 1990; Meuleau & Bourgin, 1999; Wiering & Schmidhuber, 1998). Most of these schemes calculate the uncertainty in some of the estimated quantities and add an exploration bonus based on this to a CE estimate of V_i . The first was Kaelbling’s interval estimation method. Applied to bandit tasks this selects the action with the highest upper bound on an interval estimate of the immediate reward. When applied directly to Q-learning in multi-stage decision problems it uses a window or decaying trace of previous estimated Q-values to generate the upper bounds on the Q-values. This, however, means that the estimate picks up the non-stationarity in the Q-values due to their initial bias. In addition the local exploration bonus is only combined with the estimated Q-values for action selection purposes. The bonus is not propagated to predecessor states and thus the resulting measure is local rather than distal. Meuleau and Bourgin (1999) created a distal IE measure by combining the local IE bonus δ_i^a with the reward so that it is propagated to predecessor states in the estimated model:

$$\xi_i^a = \delta_i^a (1 - \gamma) + \sum_j \bar{p}_{ij}^a (r_{ij}^a + \gamma \max_b (\xi_j^b)) \quad (5)$$

where δ_i^a is the local bonus, $(1 - \gamma)$ is the scaling factor based on the discount rate, and ξ_i^a is the exploratory value of taking action a in state i . The agent then follows a policy which is greedy with respect to ξ_i^a . One version of this algorithm (variance-based) also uses a window of previous Q values to calculate the local exploration bonus; while their worst case method uses an upper bound on the underlying variance in the return. Some form of asynchronous real time dynamic programming (ARTDP) is used to adjust the exploration value function on-line. Meuleau’s methods have been shown to outperform most current exploration techniques on a variety of tasks.

Wiering and Schmidhuber (1998) extended the interval estimation concept in a different way. Rather than estimating the variance in the Q-values directly and using this to supply an exploration bonus, we can apply the optimism heuristic to the estimated transition function \hat{P} . For each state action pair the upper bound of the $(1 - \alpha)100\%$ confidence interval is calculated for the transition probability leading to the successor state

Optimal Model Selection

Initialise $V_k, m_{ij}^a, \forall i, j \in \mathcal{S} \cup k$ and $\forall a \in \mathcal{A}$
 $\xi_i^a = \gamma V_k, \forall i \in \mathcal{S}$ and $\forall a \in \mathcal{A}$ where $m_{ik}^a > 0$
 $\xi_k^a = V_k, \forall a \in \mathcal{A}$
 repeat
 observe x
 select $a = \arg \max_b \{\xi_x^b\}$ breaking ties randomly
 observe the transition $x \xrightarrow{a} y$
 $m_{xy}^a = m_{xy}^a + 1$
 until your ARTDP algorithm stops
 choose i
 for each action b
 find $\vec{p}_{opt,i}^b$ using OMS-s or OMS-f
 update ξ_i^b using Eq. 6

Figure 1. The OMS algorithm main loop.

OMS-s

for i, a construct $\vec{p}_{opt,i}^a$:
 $p_{opt,ik}^a = \text{upperbound}(\text{Beta}, m_{ik}^a, \sum_{j \neq k} m_{ij}^a, \alpha)$
 $p_{opt,ij}^a = \frac{1 - p_{opt,ik}^a}{1 - \bar{p}_{ik}^a} \bar{p}_{ij}^a, \forall j \neq k$
 where $\bar{p}_{ij}^a = \frac{m_{ij}^a}{\sum_{x \in \mathcal{S} \cup k} m_{ix}^a}$

Figure 2. Simple optimistic model selection.

with the highest estimated value. The other transition probabilities are renormalised, and ARTDP is applied to the optimistic MDP generated. One drawback to this method is that they use a Gaussian density to model the uncertainty about each transition probability. Consequently the sample sizes for each transition have to be large before that assumption is justified. Because of this they initially employ a distal counter-based exploration method to acquire a good estimated model, and then use model-based interval estimation (MBIE) to bias exploration to the most useful (highly rewarding) parts of the state space. The algorithm switches to MBIE when the changes in the value function become small and so loses the advantage of MBIE during the early stages of exploration.

3. Optimistic Model Selection

We integrate the idea of MBIE with the Bayesian view of exploration by selecting an optimistic model P_{opt} from \mathcal{P} using probability intervals calculated based on $f(P|M)$. Since the Dirichlet is the natural conjugate density for sampling from a multinomial distribution, this allows us to correctly incorporate prior knowledge

OMS-f

for i, a construct the model $\vec{p}_{opt,i}^a$:
 order the u successors of i, a that are in $\mathcal{S} \cup k$ to give (j_1, j_2, \dots, j_u) such that $V_{j_1} \geq V_{j_2} \geq \dots \geq V_{j_u}$
 where $V_j = \max_b \{\xi_j^b\}$
 $p_{\uparrow,ix}^a = \text{upperbound}(\text{Beta}, m_{ix}^a, \sum_{y \neq x} m_{iy}^a, \alpha), \forall x$
 $p_{\downarrow,ix}^a = \text{lowerbound}(\text{Beta}, m_{ix}^a, \sum_{y \neq x} m_{iy}^a, \alpha), \forall x$
 set s to be as large as possible while
 $p_{\downarrow,ij_s}^a \leq 1 - (\sum_{p < s} p_{\uparrow,ij_p}^a + \sum_{q > s} p_{\downarrow,ij_q}^a) \leq p_{\uparrow,ij_s}^a$
 set $p_{opt,ij_p}^a = p_{\uparrow,ij_p}^a, \forall p < s$
 set $p_{opt,ij_q}^a = p_{\downarrow,ij_q}^a, \forall q > s$
 set $p_{opt,ij_s}^a = 1 - \sum_{j \neq j_s} p_{opt,ij}^a$

Figure 3. Full optimistic model selection.

about the transition function, and to explore using OMS from the outset. The main problem is how to choose the prior matrix M in the case where we have very little knowledge about P . If we do not know which transitions are possible we can either assign a prior to all possible transitions; or we can assign a prior to some subset. We take the second case to the limit by using a single additional terminal state k to represent possible unobserved transitions (Kearns & Singh, 1991). By making this state highly rewarding (Moore & Atkeson, 1993) we can also induce a distal exploration value function that will drive the learner toward novel state action pairs. If the value V_k of this absorbing state k is an upper bound on the true value function, then any initial model which has $p_{ik}^a = 1$ for all i, a will be an upper bound on the value function in all states. It will also be the most optimistic model given V_k . If we have little prior knowledge of the MDP then the parameter matrix M is initially all zero except for a single hypothesised transition to the terminal state k from every state action pair i, a , the prior for which is m_{ik}^a . Each time t the agent selects an action a in state i that maximises ξ_i^a and observes the transition $i \xrightarrow{a} j$. It then updates the parameter matrix M in the standard way. Because of our degenerate prior, each time a novel transition is observed this update is not Bayesian since observations are incorporated for previously excluded hypotheses. Subsequent updates follow Bayes rule. How exactly should we select an optimistic model? We suggest two ways, which we term simple OMS (Figure 2) and full OMS (Figure 3). In simple OMS we are optimistic only about the hypothesised transition to the terminal state. In full OMS we can be optimistic about transitions to other states too. In simple OMS once we are given the new information, we re-calculate the upper bound of the $(1 - \alpha)$ probability interval for the

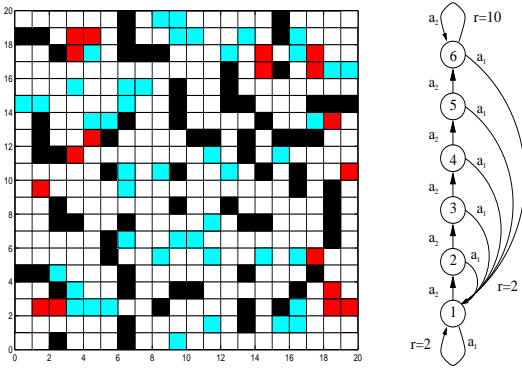


Figure 4. The task environments. (a) A stochastic maze. Walls are marked in black, and penalty fields of -4 and -1 in dark and light grey respectively. (b) Meuleau’s deceptive task. In Task 3 the positive rewards in the maze indexed by (x,y) coordinates were as follows: $r(11,1)=10$, $r(19,3)=8$, $r(19,4)=9$, $r(1,7)=7$, $r(5,20)=5$.

transition probability to the state k for each state action pair, denoted $p_{opt,ik}^a(M)$. The marginal density required for this computation is simply a Beta density, always following $\text{Beta}(m_{ik}^a, \sum_{j \neq k} m_{ij}^a)$. Since we know the single prior parameter m_{ik}^a and also that all other parameters are integer we can calculate the upper bounds for a suitable parameter set off-line using a method of successive approximations. To create a lookup table for this function for a reasonable range of values of m_{ij}^a , we take advantage of the fact that the function is smooth and changes slowly at high values, calculating it for logarithmically spaced points. We then use linear interpolation in order to provide an estimate of the upper bound of the $(1 - \alpha)$ probability interval. This is accurate in our implementation to 4 decimal places. The other probabilities are then renormalised to give an optimistic one step transition model from i, a . Applied to all states the result is an optimistic MDP P_{opt} . An ARTDP method can then be applied to P_{opt} to revise our estimate of the value function (Figure 1). The relevant Bellman equation is:

$$\xi_i^a(M) = \sum_j p_{opt,ij}^a(M) (r_{ij}^a + \gamma \max_{a'} \{\xi_j^{a'}(M)\}) \quad (6)$$

where $p_{opt,ij}^a(M)$ are the transition probabilities according to P_{opt} . The agent then selects the action with the highest optimistic value ξ_i^a .

Simple OMS can be seen as a relation of Kearns and Singh’s E^3 algorithm in which the learner chooses either to identify the model by taking actions that drive it toward the unknown state set, or to exploit within the set of known states. In our algorithm as soon as a state action pair is tried it is considered known, and can be used in exploitation if it is appealing enough.

Table 1. Algorithm Parameters

Parameter Settings	
variance based IEDP+	window length = 30 task 1 $\delta_1 = 10, 50, 100, 300, 400, 500, 1000$ task 2 $\delta_1 = 46, 60, 100, 177, 200, 400, 1000$ task 3 $\delta_1 = 0, 20, 40, 80, 100, 200, 400, 1000$
Wiering’s MBIE	switching parameter $\eta = 2^c$ task 1 $c = 3, 2, 1, 0, -1, -2, -3, -4, -5, -6$ $K_c = 50, \alpha = 0.05$ task 2 $c = 0, -1, -2, -3, -4, -5, -6$ task 3 $c = 3, 2, 1, 0, -1, -2, -3, -4, -5, -6$
OMS simple and full	$m_{ik}^a = 2^d \alpha = 0.05$ task 1 $d = 0, -1, -2, -3, -4, -5, -6, -7$ $V_k = 316, 400, 700, 1000$ task 2 $d = -3, -4, -5, -6, -7, V_k = 100$ task 3 (s) $d = 1, 0, -1, \dots -7$ task 3 (f) $d = -1, -2, -3, \dots -8$

In the full OMS algorithm the model can be optimistic about the transition probabilities for any of the successors of i, a . To achieve this we employ the idea of bounded parameter MDPs (Givan et al., 1997). Rather than perform interval value iteration, we compute only the optimistic value function. Given state action pair i, a we order its successors by the current estimate of the value function, in descending order. We then calculate the lower and upper bounds of the $(1 - \alpha)$ probability interval for each transition. An optimistic transition function is then constructed by sending as much probability mass as possible to the states early in the ordering, while keeping all probabilities within their lower and upper bounds. The state action pair i, a is then backed up using the optimistic one-step transition function that results. Full OMS can be seen as an extension of Wiering and Schmidhuber’s method which uses a more appealing density to represent uncertainty about the model; utilises this density in exploration control from the outset; and takes account of all successors in calculating the optimistic model.

If V_k is an upper bound on the value function then the transition $i \xrightarrow{a} k$ will always have the most optimistic estimate. Since the ordering of the successors j may change as we perform asynchronous back-ups and the value function changes, the sorting needs to be performed every time a state action pair is backed up. This optimistic estimate can be calculated using any form of ARTDP. We have implemented it using Wiering and Schmidhuber’s version of prioritised sweeping. In each algorithm we need a rule to set an upper bound on the value function. Following Meuleau we assume that information about the reward function is available to us. The worst case upper bound is $r_{max}/(1 - \gamma)$. If we know more about a process then it may be possible to derive a tighter upper bound. If we also assume

Table 2. Results for Measure 1, $\sum_{t=T_1}^{T_2} r_t$. For Task 1 $T_1 = 1$, $T_2 = 5000$; Task 2 $T_1 = 1$, $T_2 = 25000$; Task 3 $T_1 = 5001$, $T_2 = 25000$

Algorithm	Task 1	Task 2	Task 3
IEDP+ (vb)	14450 $\delta_1 = 1000$	60357 $\delta_1 = 1000$	34761 $\delta_1 = 1000$
IEDP+ (wc)	14100	-6915	—
Wiering's MBIE	12300 $c = 0$	58562 $c = -1$	23173 $c = 2$
OMS-s	14900 $V_k = 316$ $d = -4$	58465 $d = -6$	35543 $d = -5$
OMS-f	14950 $V_k = 316$ $d = -1$	57200 $d = -6$	36252 $d = -5$
sOMS-f (with settling)	—	—	38027 $d = -4$

Table 3. Results for Measure 3, (policy quality).

Algorithm	Task 1	Task 2	Task 3
	% runs π^* found	$1 - V^{\pi_f}(i_0)/V^*(i_0)$	
IEDP+ (vb)	94 $\delta_1 = 1000$	0.0468 $\delta_1 = 1000$	0.0429 $\delta_1 = 100$
IEDP+ (wc)	91	0.0532	—
Wiering's MBIE	100 $c = -6$	0.0105 $c = -3$	0.0397 $c = -1$
OMS-s	100 $V_k = 1000$ $d = (-6,0)$	0.0116 $d = -4$	0.0211 $d = -1$
OMS-f	100 $V_k = 1000$ $d = (-6,0)$	0.0105 $d = -4$	0.0211 $d = 1$
sOMS-f (with settling)	—	—	0.0527 $d = -4$

that we typically know whether states are terminal or not then if the highest value is in a terminal state we simply pick r_{max} .

4. Empirical Study

We compared simple and full OMS with Wiering's MBIE algorithm and Meuleau's variance based and worst case IEDP+ algorithms. The environments were Meuleau's deceptive MDP (Task 1 – Figure 3(b)), and a 400 state MDP maze (Tasks 2 and 3 – Figure 3(a)). In the maze the starting state is in the centre of the maze (x,y = 11,10). There are four actions (N,S,E,W) and transitions have Pr(0.8) of succeeding, 0.08 of carrying the agent laterally to the intended direction, and 0.04 of carrying the agent one step in the opposite direction. Reward in this environment is a deterministic function of state. In Task 2 the four corners are terminal states, the top right corner generating a reward of 100, and the other three rewards of 50 each.

Table 4. Results for Measure 2, $\sum_{t=T_1}^{T_2} \gamma^{t-1} r_t$. T_1 and T_2 are the same as for Measure 1.

Algorithm	Task 1	Task 2	Task 3
IEDP+ (vb)	186 $\delta_1 = 500$	-2.81 $\delta_1 = 60$	5.41 $\delta_1 = 100$
IEDP+ (wc)	187	—	—
W-MBIE	154 $c = 2$	-2.88 $c = -1$	4.32 $c = 3$
OMS-s	216 $V_k = 316$ $d = -4$	-2.91 $d = -5$	3.12 $d = -4$
OMS-f	259 $V_k = 316$ $d = -7$	-2.84 $d = -5$	6.91 $d = -4$
sOMS-f (with settling)	—	—	72.8 $d = -3$

The maze is filled with penalty fields (-4 or -1) and walls. The transitions for actions that lead to walls are redirected into the state in which the action was taken. In Task 3 we altered the reward structure after 5000 steps to test the algorithms' ability to perform task transfer. The penalty fields and terminal states were retained, but the rewards in the terminal states were set to 0. Instead positive rewards were allocated to states as described in the caption of Figure 3. In Task 1 $\gamma = 0.99$ and in the maze experiments 0.95. All algorithms were tested on the first two tasks. In Task 1 each run of an algorithm was 5000 time steps. In Task 2 each run consisted of approximately 25000 time steps, and possibly of many trials. The last trial in each run was allowed to terminate even if it meant the total run length exceeded 25000 steps. In the task transfer experiment the algorithms were run with the initial reward structure for 5000 steps, and then for another 5000 steps given the new reward function.

All algorithms were optimised across a parameter set. Each algorithm was tested for 100 runs of each setting on each task. The parameters were as shown in Table 1. In Task 1 V_k was set to a range of values to test the drop in performance as the upper bound on V increases. In Task 2, and the first 5000 ticks of Task 3 all algorithms were given the benefit of the information that the high rewarding states were terminal. In Task 3, after 5000 ticks V_k was reset using the worst case heuristic. In the case of OMS-f we also tested the performance of the algorithm when we conduct asynchronous DP between ticks 5000 and 5001 until the value function settles, rather than just performing 80 backups. We refer to this as OMS-f with settling. All algorithms were run using Wiering and Schmidhuber's version of prioritized sweeping (1998), with the threshold for the priority queue $\epsilon = 0.001$, and the maximum number of backups per step $U_{max} = 80$.

5. Results

We measured performance according to three criteria. The first is the total reward generated over the length of a run, averaged over all 100 runs. This is used following Meuleau (1999) and others. The second measure is the discounted cumulative reward averaged over all runs. The third measure assesses the expected regret of a greedy policy generated from the final agent model. In order to find the policy for each agent we extract the maximum likelihood transition model, and apply DP, to obtain a greedy policy that breaks ties between optimal actions randomly. The expected regret of this policy, π_f , is calculated using the known MDP. Finally we present the regret in the starting state i_0 as a proportion of the value of an optimal policy in that state. This measures performance on exploration criterion (ii). We did this for Tasks 2 and 3 only. In Task 1 the policy was usually optimal, so we present the % of runs for which this was the case. Due to lack of space we present the full results only for the optimised parameters for each algorithm on each criterion in Tables 2, 3, and 4. In these use of (a, b) specifies a set of values observed in that interval. We carried out t-tests for all criteria except (ii) on Task 1 where we used Fisher’s exact test. Partial order dominance graphs for all the significance tests are shown in Figure 6. Graphs of the performance across the parameter set for all algorithms on the cumulative reward and policy quality measures are shown in Figure 6. We do not show those for the discounted cumulative reward criterion due to lack of space.

6. Discussion

When all algorithms are optimised, the OMS algorithms generally outperform the others on Tasks 1 and 3 on all measures, and on Task 2 on policy quality. Variance based IEDP+ is the best algorithm on Task 2 in terms of the cumulative reward generated. Most of these differences can be seen to be statistically significant (Figure 6). On the discounted cumulative reward criterion on Task 2 the performances range between roughly -3 and -2 over all algorithms and parameter settings, and those differences are not statistically significant. This is not surprising since on such a sparse reward task you can do little else in the initial stages other than avoid penalty fields you encountered. In Task 1 both simple and full OMS generate more reward and find better policies than any of the other three algorithms. IEDP+ never finds the correct policy for all 100 runs, whatever the parameter setting. Wiering’s MBIE does find the best policy consistently when optimised, but sacrifices considerable rewards to

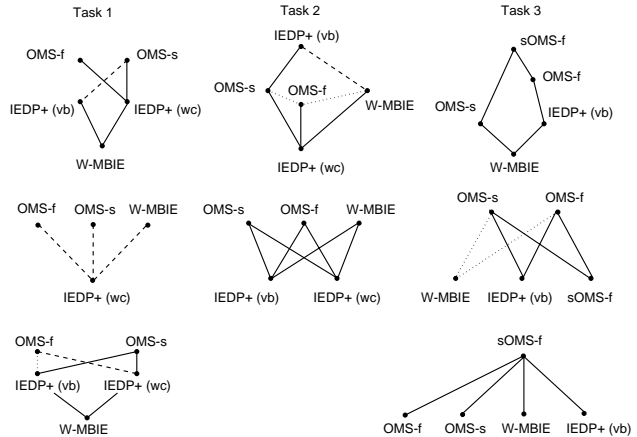


Figure 5. Partial Order Dominance for optimised parameters for Tasks 1-3. Solid lines indicate significance at 1% or higher; dashed at 5%, and dotted at 10%. The upper row shows the ordering for cumulative reward, the middle row for policy quality, and the bottom row for discounted cumulative reward. All significances were corrected according to the number of pairs in order to keep the probability of a Type I error low.

do so. Both OMS algorithms find the best policy consistently across a broad range of settings of m_{ik}^a for all values V_k of the absorbing state. If V_k is a tight upper bound on V then they do so without sacrificing reward generated. If V_k is a loose upper bound (here 1000 was the worst case estimate) then they only generate high rewards for some values of m_{ik}^a : cumulative reward falls steadily as m_{ik}^a increases (see Figure 6). As V_k increases the effect becomes more pronounced. This makes sense since m_{ik}^a governs how slowly $p_{opt,ik}^a$ falls. It reflects optimism about the likelihood of reaching k directly, and thereby governs persistence. V_k reflects our optimism about what we will find if we get there. In Task 2 variance based IEDP+ outperforms all other algorithms on the cumulative reward criterion. In addition on this task, it performs consistently well across the parameter space, whereas all other algorithms vary significantly as their parameters are altered. The other algorithms have similar performance to each other when optimised, except for worst case IEDP+, which performs very badly indeed. IEDP+ fails to find policies as close to optimal as the other algorithms, always generating policies some 5% worse than optimal. OMS and Wiering’s MBIE both generate policies about 1% below optimal, and sacrifice considerable rewards to do so. Both OMS algorithms fail to do well on both criteria at once. Either high rewards are generated, or good policies are discovered. In order to generate high rewards, OMS generates policies some 9% below optimal. This contrasts with Wiering’s algorithm, which usually generated policies around 3% below optimal.

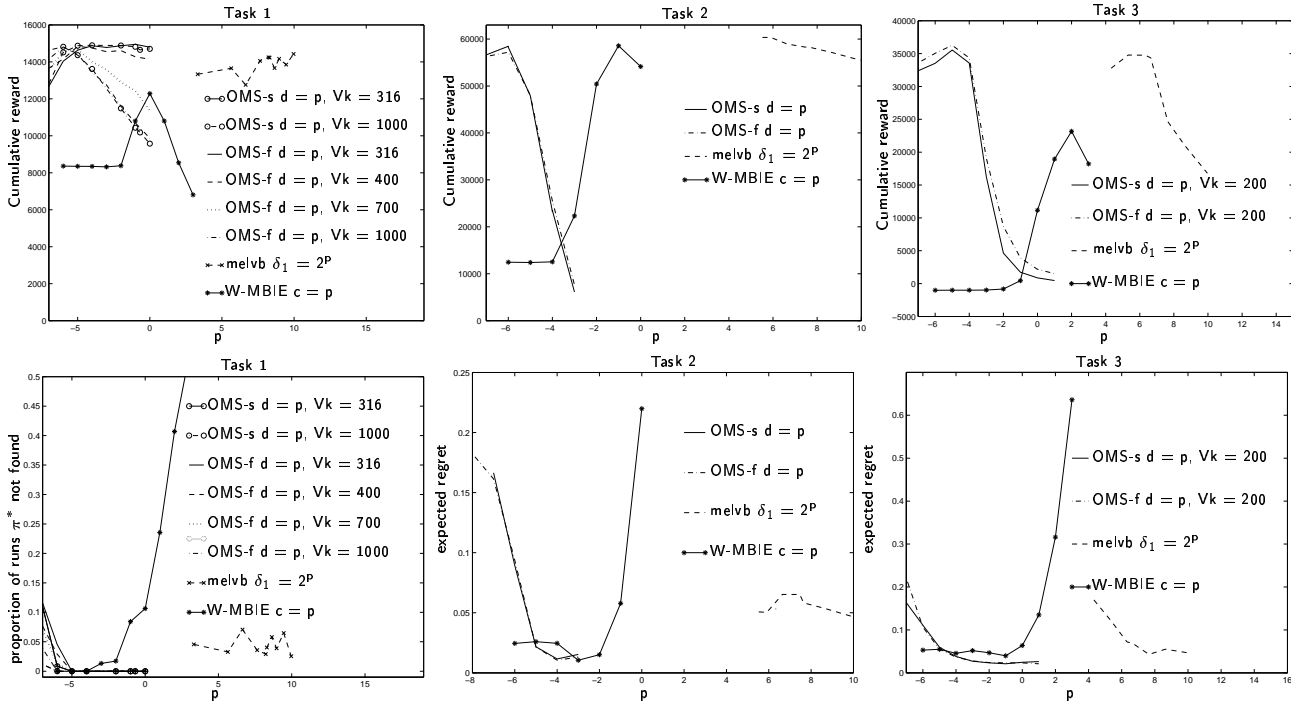


Figure 6. Change in the performance of algorithms over the parameter space for Tasks 1-3. To put all algorithms on the same graph we use a unifying parameter p . Expected regret refers to the value of $1 - \frac{V^{\pi^* f}(i_0)}{V^*(i_0)}$ calculated as discussed in Section 5.

Finally in Task 3 both versions of OMS outperform both other algorithms on policy quality and cumulative reward. Here the repeated availability of the reward causes the agent to need considerable impetus to explore the states around so as to improve the policy sufficiently. OMS is therefore optimised in terms of finding good policies for a higher value of m_{ik}^a than in Tasks 1 and 2. Wiering’s algorithm carries out counter-based exploration until the value function has settled. Thus its performance in task transfer will depend heavily on the ratio of computation available per step to computation required to stabilise V . Both Wiering’s MBIE and OMS have to sacrifice almost all the available reward in order to improve the final policy to within 2% of optimality. IEDP+ never generates policies within an average of 4% of optimality, and often does worse. In this task no algorithms performed well on either criterion across all parameter values. It is not clear why IEDP+ varies significantly on Task 3, while doing consistently well on the cumulative reward criterion across the parameter range in Tasks 1 and 2. On the discounted cumulative reward criterion there are small differences, but most of these are not statistically significant. However, W-MBIE and OMS perform much better if settling is allowed (we only present results for settling OMS-f here, but expect W-MBIE

to show similar performance). Here the performance of OMS-f rises from an average discounted cumulative reward of 6.9 to 72.8 with settling. There is a rise in the cumulative reward as well, and a commensurate worsening of policy quality. Meuleau’s variance based algorithms can never take advantage of prior knowledge in this way because they require new experiences to update the exploration value function. Thus if sufficient computation is available, placing uncertainty in the model rather than the value function, gives a considerable advantage in task transfer. This is in a sense the strongest result, since it is in problems involving knowledge transfer that exploration control can give the greatest benefits.

7. Parameter Selection

There are several points to note when judging these algorithms on their performance across the parameter range; and in selecting parameter values when applying them to problems. The first is that all algorithms vary significantly in performance across the parameter range for at least one task. The second is that the Bayesian algorithms’ parameters are precisely priors that reflect the experimenter’s beliefs about the task. The Bayesian algorithms perform poorly when the pri-

ors given by the experimenter are misleading. This is only to be expected, since the algorithms are acting according to the beliefs they have been given, and any exploration method should alter its exploration policy as the view of the world it is given alters. In terms of the the number of trials of each state action pair weak priors are in fact rapidly washed out; it is simply a large number of transitions, 6400 in Task 2 for example, that requires a long period of exploration. Third, the technique employed here of using an additional state to represent priors for unknown transitions compactly and approximately extends the range of problems to which Bayesian OMS can be applied. The correct Bayesian formulation requires separate priors for all possible transitions. It would be perfectly possible to apply OMS in that framework. Fourth, it is sensible to extend the technique where required, e.g. by using a prior over V_k ; and thereby using knowledge of the existence of local structure in the MDP, by creating a prior for V_{ik}^a based on the estimated values of nearby states. Indeed it should be possible to devise an algorithm which estimates the degree of local structure as it learns, and uses this to modify its density over the model space.

8. Conclusion and Future Work

We have presented two new heuristic algorithms which outperform two of the leading model-based explorers when all algorithms are optimised. The performance of all the algorithms varies significantly across the parameter set for one or more tasks. A key question is therefore how other limited problem knowledge, e.g. the likely density of connections, or knowledge of the existence of local structure, can be used to guide parameter selection. Finally we argue that the major role of exploration control will be in task transfer. Therefore a clear and tractable method for handling prior knowledge, and uncertainty about process models is important. This is where Bayesian approaches come into their own. Given the intractability of reasoning using the precise Bayesian formulation, OMS based on Dirichlet densities provides a tractable alternative. There are several extensions to be made to OMS. We are currently extending the method to the construction of optimistic multi-time models in hierarchical reinforcement learners in order to guide exploration over options in semi-MDPs. We are also working to generalise the technique to stochastic process models with factored representations.

References

- Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.
- Dearden, R., Friedman, N., & Andre, D. (1999). Model-based Bayesian exploration. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)* (pp. 150–159). San Francisco, CA: Morgan Kaufmann Publishers.
- Fiechter, C. (1997). Expected mistake bound model for on-line reinforcement learning. *Proceedings of the 14th International Conference on Machine Learning* (pp. 116–124). Morgan Kaufmann.
- Gittins, J. (1989). *Multi-armed bandit allocation indices*. Interscience Series in Systems and Optimization. John Wiley & Sons.
- Givan, R., Leach, S., & Dean, T. (1997). Bounded parameter Markov decision processes. *Recent Advances in AI Planning: 4th European Conference on Planning*. Springer Verlag.
- Kaelbling, L. P. (1990). *Learning in embedded systems*. Doctoral dissertation, Dept. of Computer Science, Stanford.
- Kearns, M., & Singh, S. (1991). Near-optimal reinforcement learning in polynomial time. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 260–268). Morgan Kaufmann.
- Martin, J. (1967). *Bayesian decision problems and Markov chains*. New York: Wiley.
- Meuleau, N., & Bourgin, P. (1999). Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35, 117–154.
- Moore, A. W., & Atkeson, C. G. (1993). Prioritised sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13, 103–130.
- Strens, M. (2000). A Bayesian framework for reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 943–950). Morgan Kaufmann.
- Wiering, M., & Schmidhuber, J. (1998). Efficient model-based exploration. *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*.
- Wyatt, J. (1997). *Exploration and inference in learning from reinforcement*. Ph.D. thesis, University of Edinburgh, Dept. of Artificial Intelligence, Edinburgh University.