The Many Faces of Optimism: a Unifying Approach

István Szita   SZITYU@GMAIL.COM
András Lőrincz   ANDRAS.LORINCZ@ELTE.HU
Eötvös Loránd University, 1117 Pázmány P. sétány 1/C, Budapest, Hungary

Abstract

The exploration-exploitation dilemma has been an intriguing and unsolved problem within the framework of reinforcement learning. "Optimism in the face of uncertainty" and model building play central roles in advanced exploration methods. Here, we integrate several concepts and obtain a fast and simple algorithm. We show that the proposed algorithm finds a near-optimal policy in polynomial time, and give experimental evidence that it is robust and efficient compared to its ascendants.

1. Introduction

Reinforcement learning (RL) is the art of maximizing long-term rewards in a stochastic, unknown environment. In the construction of RL algorithms, the choice of exploration strategy is of central significance.

We shall examine the problem of exploration in the Markov decision process (MDP) framework. While simple methods like ϵ-greedy and Boltzmann exploration are commonly used, it is known that their behavior can be extremely poor (Koenig & Simmons, 1993). Recently, a number of efficient exploration algorithms have been published, and for some of them, formal proofs of efficiency also exist. We review these methods in Section 2. By combining ideas from several sources, we construct a new algorithm for efficient exploration. The new algorithm, *optimistic initial model* (OIM), is described in Section 3. In Section 4, we show that many of the advanced algorithms, including ours, can be treated in a unified way. We use this fact to sketch a proof that OIM finds a near-optimal policy in polynomial time with high probability. Section 5 provides experimental comparison between OIM and

a number of other methods on some benchmark problems. Our results are summarized in Section 6. In the rest of this section, we review the necessary preliminaries, Markov decision processes and the exploration task.

1.1. Markov Decision Processes (MDPs)

Markov decision processes are the standard framework for RL, and the basis of numerous extensions (like continuous MDPs, partially observable MDPs or factored MDPs). An MDP is characterized by a quintuple $(X, A, \mathcal{R}, P, \gamma)$, where X is a finite set of states; A is a finite set of possible actions; $\mathcal{R} : X \times A \times X \to \mathcal{P}_\mathbb{R}$ is the reward distribution, $R(x, a, y)$ denotes the mean value of $\mathcal{R}(x, a, y)$, $P : X \times A \times X \to [0, 1]$ is the transition function; and finally, $\gamma \in [0, 1)$ is the discount rate on future rewards. We shall assume that all rewards are nonnegative and bounded from above by $R_{\max}^0$.

A (stationary) policy of the agent is a mapping $\pi : X \times A \to [0, 1]$. For any $x_0 \in X$, the policy of the agent and the parameters of the MDP determine a stochastic process experienced by the agent through the instantiation $x_0, a_0, r_0, x_1, a_1, r_1, \ldots, x_t, a_t, r_t, \ldots$

The goal is to find a policy that maximizes the expected value of the discounted total reward. Let us define the state-action value function (value function for short) of π as $Q^\pi(x, a) := E\left(\sum_{t=0}^\infty \gamma^t r_t \,\middle|\, x = x_0, a = a_0\right)$ and the optimal value function as

$$Q^*(x, a) := \max_\pi Q^\pi(x, a)$$

for each $(x, a) \in X \times A$. Let the greedy action at x w.r.t. value function Q be $a_x^Q := \arg\max_a Q(x, a)$. The greedy policy of Q deterministically takes the greedy action in each state. It is well-known that the greedy policy of Q^* is an optimal policy and Q^* satisfies the Bellman equations:

$$Q^*(x, a) = \sum_y P(x, a, y)\Big(R(x, a, y) + \gamma Q^*(y, a_y^{Q^*})\Big).$$

1.2. The Exploration Problem

In the classical reinforcement learning setting, it is assumed that the environment can be modelled as an MDP, but its parameters (that is, P and R) are unknown to the agent, and she has to collect information by interacting with the environment. If too little time is spent with the exploration of the environment, the agent will get stuck with a suboptimal policy, without knowing that there exists a better one. On the other hand, the agent should not spend too much time visiting areas with low rewards and/or accurately known parameters.

What is the optimal balance between exploring and exploiting the acquired knowledge and how could the agent concentrate her exploration efforts? These questions are central for RL. It is known that the optimal exploration policy in an MDP is non-Markovian, and can be computed only for very simple tasks like k-armed bandit problems.

2. Related Literature

Here we give a short review about some of the most important exploration methods and their properties.

2.1. ϵ-greedy and Boltzmann Exploration

The most popular exploration method is ϵ-greedy action selection. The method works without a model, only an approximation of the action value function $Q(x, a)$ is needed. The agent in state x selects the greedy action a_x^Q or an explorative move with a random action with probabilities $1 - \epsilon$ and ϵ, respectively. Sooner or later, all paths with nonzero probability will have been visited many times, so, a suitable learning algorithm can learn to choose the optimal path. It is known, for example, that Q-learning with nonzero exploration converges to the optimal value function with probability 1 (Littman & Szepesvári, 1996), and so does SARSA (Singh et al., 2000), if the exploration rate diminishes according to an appropriate schedule.

Boltzmann exploration selects actions as follows: the probability of choosing action a is $\dfrac{\exp\big(Q(s,a)/T\big)}{\sum_{a' \in A} \exp\big(Q(s,a')/T\big)}$, where 'temperature' $T \, (>0)$ regulates the amount of explorative actions. Convergence results of the ϵ-greedy method carry through to this case.

Unfortunately, for the ϵ-greedy and the Boltzmann method, exploration time may scale exponentially in the number of states (Koenig & Simmons, 1993).

2.2. Optimistic Initial Values (OIV)

One may boost exploration with a simple trick: the initial value of each state action pair can be set to some overwhelmingly high number. If a state x is visited often, then its estimated value will become more exact, and therefore, lower. Thus, the agent will try to reach the more rarely visited areas, where the estimated state values are still high. This method, called 'exploring starts' or 'optimistic initial values', is a popular exploration heuristic (Sutton & Barto, 1998), sometimes combined with others, e.g., the ϵ-greedy exploration method. Recently, Even-Dar and Mansour (2001) gave theoretical justification for the method: they proved that if the optimistic initial values are sufficiently high, Q-learning converges to a near-optimal solution. One apparent disadvantage of OIV is that if initial estimations are too high, then it takes a long to fix them.

2.3. Bayesian Methods

We may assume that the MDP (with the unknown values of P and R) is drawn from a parameterized distribution $\mathcal{M}_0$. From the collected experience and the prior distribution $\mathcal{M}_0$, we can calculate successive posterior distributions $\mathcal{M}_t, t = 1, 2, \ldots$ by Bayes' rule. Furthermore, we can calculate (at least in principle) the policy that minimizes the uncertainty of the parameters (Strens, 2000). Dearden (2000) approximates the distribution of state values directly. Exact computation of the optimal exploration policy is infeasible and Bayesian methods are computationally demanding even with simplifying assumptions about the distributions, e.g., the independencies of certain parameters.

2.4. Confidence Interval Estimation

Confidence interval estimation algorithms are between Bayesian exploration and OIV. It assumes that each state value is drawn from an independent Gaussian distribution and it computes the confidence interval of the state values. The agent chooses the action with the highest upper confidence bound. Initially, all confidence intervals are very wide, and shrink gradually towards the true state values. Therefore, the behavior of the technique is similar to OIV. The IEQL+ method of Meuleau and Bourgine (1999) directly estimates confidence intervals of Q-values, while Wiering and Schmidhuber (1998) calculate confidence intervals for P and R, and obtain Q-value bounds indirectly. Strehl and Littman (2006) improve the method and prove a polynomial-time convergence bound. Both algorithms are called *model-based interval estimation*. To avoid

confusion, we will refer to them as MBIE(WS) and MBIE(SL).

Auer and Ortner (2006) give a confidence interval-based algorithm, for which the online regret is only logarithmic in the number of steps taken.

2.5. Exploration Bonus Methods

The agent can be directed towards less-known parts of the state space by increasing the value of 'interesting' states artificially with bonuses. States can be interesting given their *frequency, recency, error*, etc. (Meuleau & Bourgine, 1999; Wiering & Schmidhuber, 1998).

The balance of exploration and exploitation is usually set by a scaling factor κ, so that the total immediate reward of the agent at time t is $r_t + \kappa \cdot b_t(x_t, a_t, x_{t+1})$, where b_t is one of the above listed bonuses. The bonuses are calculated by the agent and act as intrinsic motivating forces. Exploration bonuses for a state can vary swiftly and model-based algorithms (like prioritized sweeping or Dyna) are used for spreading the changes effectively. Alas, the weight of exploration κ needs to be annealed according to a suitable schedule.

Alternatively, the agent may learn *two* value functions separately: a regular one, Q_t^r which is based on the rewards r_t received from the environment, and an exploration value function Q_t^e which is based on the exploration bonuses. The agent's policy will be greedy with respect to their combination $Q_t^r + \kappa Q_t^e$. Then the exploration mechanism may remain the same, but several advantages appear. First of all, the changes in κ take effect *immediately*. As an example, we can immediately switch off exploration by setting κ to 0. Furthermore, Q_t^r may converge even if Q_t^e does not.

Confidence interval estimation can be phrased as an exploration bonus method: see IEQL+ (Meuleau & Bourgine, 1999) or MBIE-EB (Strehl & Littman, 2006). Even-Dar and Mansour (2001) have shown that ϵ-greedy and Boltzmann explorations can be formulated as exploration bonus methods although rewards are not propagated through the Bellman equations.

2.6. E^3 and R-max

The *Explicit explore or exploit* (E^3) algorithm of Kearns and Singh (1998) and its successor, R-max (Brafman & Tennenholtz, 2001) were the first algorithms that have polynomial time bounds for finding near-optimal policies. R-max collects statistics about transitions and rewards. When visits to a state enable high precision estimations of real transition probabilities and rewards then state is declared *known*. R-max also maintains an approximate model of the environ-

ment. Initially, the model assumes that all actions in all states lead to a (hypothetical) maximum-reward absorbing state. The model is updated each time when a state becomes known. The optimal policy of the model is either the near-optimal policy in the real environment or enters a not-yet-known state and collects new information.

3. Construction of the Algorithm

Our agent starts with a simple, but overly optimistic model. By collecting new experiences, she updates her model, which becomes more realistic. The value function is computed over the approximate model with (asynchronous) dynamic programming. The agent always chooses her action greedily w.r.t. her value function. Exploration is induced by the optimism of the model: unknown areas are believed to yield large rewards. Algorithmic components are detailed below.

Separate exploration values. Similarly to the approach of Meuleau and Bourgine (1999), we shall separate the 'true' state values from exploration values. Formally, the value function has the form

$$Q(x, a) = Q^r(x, a) + Q^e(x, a)$$

for all $(x, a) \in X \times A$, where Q^r and Q^e will summarize external and exploration rewards, respectively.

'Garden of Eden' state. Similarly to R-max, we introduce a new hypothetical 'garden of Eden' state x_E, and assume an extended state space $X' = X \cup \{x_E\}$. Once there, then, according to the inherited model, the agent remains in x_E indefinitely and receives $R_{\max}$ reward for every step, which may exceed $R_{\max}^0 =: \max_{x,a,y} R(x, a, y)$, the maximal reward of the original environment.

Model approximation. The agent builds an approximate model of the environment. For each $x, y \in X$ and $a \in A$, let $N_t(x, a)$, $N_t(x, a, y)$, and $C_t(x, a, y)$ denote the number of times when a was selected in x up to step t, the number of times when transition $x \xrightarrow{a} y$ was experienced, and the sum of external rewards for $x \xrightarrow{a} y$ transitions, respectively. With these notations, the approximate model parameters are

$$\hat{P}_t(x, a, y) = \frac{N_t(x, a, y)}{N_t(x, a)} \text{ and } \hat{R}_t(x, a, y) = \frac{C_t(x, a, y)}{N_t(x, a, y)}.$$

Suitable initializations of $N_t(x, a)$, $N_t(x, a, y)$ and $C_t(x, a, y)$ will ensure that the ratios are well-defined everywhere. The exploration rewards are defined as

$$R^e(x, a, y) := \begin{cases} R_{\max}, & \text{if } y = x_E; \\ 0, & \text{if } y \neq x_E, \end{cases}$$

for each $x, y \in X \cup \{x_E\}$, $a \in A$, and are not modified during the course of learning.

Optimistic initial model. The initial model assumes that x_E has been reached once for each state-action pairs: for each $x \in X \cup \{x_E\}$, $y \in X$ and $a \in A$,

$$N_0(x, a) = 1,$$
$$N_0(x, a, y) = 0, \qquad C_0(x, a, y) = 0.$$
$$N_0(x, a, x_E) = 1, \quad C_0(x, a, x_E) = 0.$$

Then, the optimal initial value function equals

$$Q_0(x, a) = Q_0^r(x, a) + Q_0^e(x, a) = 0 + \frac{1}{1 - \gamma} R_{\max} := V_{\max}$$

for each $(x, a) \in X' \times A$, analogously to OIV.

Dynamic programming. Both value functions can be updated using the approximate model. For each $x \in X$, let a_x be the greedy action according to the combined value function, i.e.,

$$a_x := \arg\max_{a \in A} \big(Q^r(x, a) + Q^e(x, a)\big).$$

The dynamic programming equations for the value function components are

$$Q_{t+1}^r(x, a) := \sum_{y \in X} \hat{P}_t(x, a, y) \left(\hat{R}_t(x, a, y) + \gamma Q_t^r(y, a_y) \right)$$

$$Q_{t+1}^e(x, a) := \gamma \sum_{y \in X} \hat{P}_t(x, a, y) Q_t^e(y, a_y)$$
$$+ \hat{P}_t(x, a, x_E) V_{\max}.$$

Episodic tasks can be handled as usual way; we introduce an absorbing final state with 0 external reward.

Asynchronous update. The algorithm can be online, if instead of full update sweeps over the state space updates are limited to state set L_t in the 'neighborhood' of the agent's current state. Neighborhood is restricted by computation time constraints; any asynchronous dynamic programming algorithm suffices. It is implicitly assumed that the current state is always updated, i.e., $x_t \in L_t$. In this paper, we used the improved prioritized sweeping algorithm of Wiering and Schmidhuber (1998).

Putting it all together. The method is summarized as Algorithm 1.

4. Analysis

In the first part of this section, we analyze the similarities and differences between various exploration methods, with an emphasis on OIM. Based on this analysis, we sketch the proof that OIM finds a near-optimal policy in polynomial time. Details of the proof can be found in (Szita & Lőrincz, 2008).

4.1. Relationship to Other Methods

'Optimism in the face of uncertainty' is a common point in exploration methods: the agent believes that she can obtain extra rewards by reaching the unexplored parts of the state space.

Note that as far as the combined value function Q is concerned, OIM is an asynchronous dynamic programming method augmented with model approximation.

Optimistic initial values. Apparently, OIM is the model-based extension of the OIV heuristic. Note however, that optimistic initialization of Q-values is not effective with a model: the more updates are made, the less effect the initialization has and it fully diminishes if value iteration is run until convergence. Therefore, naive combination of OIV and model construction is contradictory: the number of DP-updates should be kept low in order to save the initial boost, but it should be as high as possible in order to propagate the real rewards quickly.

OIM resolves this paradox by moving the optimism into the model. The optimal value function of the initial model is $Q_0 \equiv V_{\max}$, corresponding to OIV. However, DP updates can not, but only model updates may lower the exploration boost.

Note that we can set the initial model value as high as we like, but we do not have to wait until the initial boost diminishes, because Q^r and Q^e are separated.

R-max. The 'Garden of Eden' state x_E of OIM is identical to the fictitious max-reward absorbing state of R-MAX (and E^3). In both cases, the agent's model tells that all unexplored (x, a) pairs lead to x_E. R-MAX, however, updates the model only when the transition probabilities and rewards are known with high precision, which is only after many visits to (x, a). In contrast, OIM updates the model after each single visit, employing each bit of experience as soon as it is obtained. As a result, the approximate model can be used long before it becomes accurate.

Exploration bonus methods. The extra reward offered by the Garden of Eden state can be understood as an exploration bonus: for each visit of the pair (x, a), the agent gets the bonus $b_t(x, a) = \frac{1}{N_t(x,a)} \big(V_{\max} - Q_t(x, a)\big)$. It is insightful to contrast this formula with those of the other methods like the *frequency-based* bonus $b_t = -\alpha \cdot N_t(x, a)$ or the *error-based* bonus $b_t = \alpha \cdot \big|Q_{t+1}(x, a) - Q_t(x, a)\big|$.

Model-based interval exploration. The exploration bonus form of the MBIE method of Strehl and Littman (2005) sets $b_t = \frac{\alpha}{N_t(x,a)}$. MBIE-EB is not an ad-hoc method: the form of the bonus comes from

---

Algorithm 1 The Optimistic initial model algorithm

> **Input:** $x_0 \in X$ initial state, $\epsilon > 0$ required precision, optimism parameter $R_{\max}$
> **Model initialization:** $t := 0$; $\forall x, y \in X, \forall a \in A$:
> $N(x, a, y) := 0$, $N(x, a, x_E) := 1$, $N(x, a) := 1$, $C(x, a, y) := 0$, $Q^r(x, a) := 0$, $Q^e(x, a) := R_{\max}/(1 - \gamma)$;
> **repeat**
> $a_t := $ greedy action w.r.t. $Q^r + Q^e$; apply a_t and observe r_t and x_{t+1}
> $C(x_t, a_t, x_{t+1}) := C(x_t, a_t, x_{t+1}) + r_t$; $N(x_t, a_t, x_{t+1}) := N(x_t, a_t, x_{t+1}) + 1$; $N(x_t, a_t) := N(x_t, a_t) + 1$
> $L_t := $ list of states to be updated
>   **for each** $x \in L_t$ **do**
> $Q^r_{t+1}(x, a) := \sum_{y \in X} \hat{P}(x, a, y) \left(\hat{R}(x, a, y) + \gamma Q^r_t(y, a_y) \right)$
> $Q^e_{t+1}(x, a) := \hat{P}(x, a, x_E) R_{\max}/(1 - \gamma) + \gamma \sum_{y \in X} \hat{P}(x, a, y) Q^e_t(y, a_y)$.
>   **end for**
> $t := t + 1$
> **until** Bellman-error$> \epsilon$

---

confidence interval estimations. The comparison to MBIE-EB will be especially valuable, as it converges in polynomial-time and the proof can be transported to OIM with slight modifications.

4.2. Polynomial-time Convergence

Theorem 4.1 *There exists a constant C so that for any $\epsilon > 0$, $\delta > 0$, $\epsilon_1 := \epsilon/4$, $\epsilon_2 := \epsilon_1(1 - \gamma)$, $H := \frac{1}{1-\gamma} \ln \frac{R^0_{\max}}{\epsilon_1(1-\gamma)}$, $K := \left(\frac{2C|X|^5|A|H^4}{\delta \epsilon_1^4} \ln \frac{1}{\delta} \right)^{1/6}$, OIM converges almost surely to a near-optimal policy in polynomial time if started with $R_{\max} \geq \frac{3}{\epsilon_2(1-\gamma)} (R^0_{\max})^2 \ln(K R^0_{\max})$, that is, with probability $1 - \delta$, the number of timesteps where $Q^{\pi^{OIM}}(x_t, a_t) > Q^*(x_t, a_t) - \epsilon$ does not hold, is $O\left((\frac{|X||H|}{\epsilon_1})^5 \frac{|A|(R^0_{\max})^7}{(1-\gamma)\epsilon_1^5} \ln^2 \frac{1}{\delta} \right)$.*

For the sketch of the proof, we shall follow the technique of Kearns and Singh (2002) and Strehl and Littman (2006), and will use the shorthands [KS] and [SL] for referring to them. See (**?**) for the detailed proof with a slightly better polynomial bound.

A pair (x, a) is declared known, if it has been visited at least $m = C(\frac{|X||H|}{\epsilon_1})^4 (R^0_{\max})^6 \ln \frac{1}{\delta}$ times, with a suitable constant C. OIM preserves the optimism of the value function:

Lemma 4.2 *Let Q_t be the sequence of Q-functions generated by OIM. Then, it holds with probability $1 - \delta/2$ that for any t, $Q_t(x, a) \geq Q^*(x, a) - \epsilon_1$.*

Proof: According to [SL], with probability $1 - \delta/2$,

$$\sum_y \hat{P}_t(x, a, y) \left(\hat{R}_t(x, a, y) + \gamma V^*(y) \right) \tag{1}$$

$$- Q^*(x, a) \geq -\beta/\sqrt{N_t(x, a)},$$

where $\beta := R^0_{\max}/(1 - \gamma)\sqrt{\ln(2|X||A|m/\delta)/2} = \frac{\sqrt{3}}{1-\gamma} R^0_{\max} \sqrt{\ln(K R^0_{\max})}$.

We will show that

$$R_{\max}/(N_t(x, a)(1 - \gamma)) + \epsilon_2 \geq \beta/\sqrt{N_t(x, a)}. \tag{2}$$

For $N_t(x, a) \leq \frac{R_{\max}}{(1-\gamma)\epsilon_2}$, the first term dominates the l.h.s. and we can omit the second term (and prove the stricter inequality). If the relation is reversed, then the first term can be omitted. In both cases, we arrive at the requirement $R_{\max} \geq \frac{3}{\epsilon_2(1-\gamma)} (R^0_{\max})^2 \ln(K R^0_{\max})$, which holds by assumption.

At step t, a number of DP updates are carried out. We proceed by induction on the number of DP-updates. Initially, $Q^{(0)}(x, a) \geq Q^*(x, a) - \epsilon_1$, then $Q^{(i+1)}(x, a) =$
$\sum_y \hat{P}_t(x, a, y) \left(\hat{R}_t(x, a, y) + \gamma V^{(i)}(y) \right) + \frac{V_{\max}}{N_t(x,a)}$
$\geq \sum_y \hat{P}_t(x, a, y) \left(\hat{R}_t(x, a, y) + \gamma (V^*(y) - \epsilon_1) \right) + \frac{V_{\max}}{N_t(x,a)}$
$\geq Q^*(x, a) - \beta/\sqrt{N_t(x, a)} - \gamma\epsilon_1 + \frac{V_{\max}}{N_t(x,a)}$
$\geq Q^*(x, a) - \gamma\epsilon_1 - \epsilon_2 = Q^*(x, a) - \epsilon_1$,
where we applied (1), (2), the induction assumption and the definition of ϵ_2. $\qquad \square$

Let M denote the true (and unknown) MDP, let $\hat{M}$ be the approximate model of OIM, and define $\bar{M}$ so that it is identical to M for known pairs, and equals $\hat{M}$ for unknown pairs. The parameters of $\hat{M}$ and $\bar{M}$ are nearly identical: if (x, a) becomes known, then the local values of $\hat{P}$ and $\hat{R}$ are $O(\frac{\epsilon}{|X|HR^0_{\max}})^2$-approximations of P and R with probability $1 - \delta/2$ (Lemma 5 of [KS]). Therefore, by Lemma 4 of [KS],

$$|Q^\pi_{\hat{M}}(x, a) - Q^\pi_{\bar{M}}(x, a)| < \epsilon_1. \tag{3}$$

Define the H-step truncated value function of policy π as $Q^\pi(x, a, H) := E\left(\sum_{t=0}^H \gamma^t r_t \mid x = x_0, a = a_0 \right)$. According to [KS] Lemma 2, $H = \frac{1}{1-\gamma} \ln \frac{R^0_{\max}}{\epsilon_1(1-\gamma)}$ is an

ϵ_1-horizon time, i.e., $|Q_M^\pi(x, a, H) - Q_M^\pi(x, a)| < \epsilon_1$ for any (x, a), π and any MDP M with discount factor γ.

Consider a state-action pair (x_1, a_1) and a H-step long trajectory generated by π. Let A_M be the event that an unknown pair (x, a) is encountered along the trajectory. Then, by Lemma 3 of [SL],

$$Q_M^\pi(x_1, a_1) \geq Q_{\tilde{M}}^\pi(x_1, a_1) - V_{\max} \Pr(A_M). \qquad (4)$$

To conclude the proof, we separate two cases (following the line of thoughts of Theorem 1 in [SL]). In the first case, an exploration step will occur with high probability: Let $V_{\max}^0 := R_{\max}^0/(1 - \gamma)$. Suppose that $\Pr(A_M) > \epsilon_1/V_{\max}^0$, that is, an unknown pair (x, a) is visited in H steps with high probability. This can happen at most $m|X||A|$ times, so by the Hoeffding-Azuma bound, with probability $1 - \delta/2$, all (x, a) will become known after $O(\frac{m|X||A|HV_{\max}^0}{\epsilon_1} \ln \frac{1}{\delta})$ exploration steps.

On the other hand, if $\Pr(A_M) \leq \epsilon_1/V_{\max}^0$, then the policy is near-optimal with probability $1 - \delta$:

$$Q_M^{\pi^{OIM}}(x_1, a_1) \geq Q_M^{\pi^{OIM}}(x_1, a_1, H)$$
$$\geq Q_{\tilde{M}}^{\pi^{OIM}}(x_1, a_1, H) - V_{\max}^0 \Pr(A_M)$$
$$\geq Q_{\tilde{M}}^{\pi^{OIM}}(x_1, a_1, H) - \epsilon_1 \geq Q_{\tilde{M}}^{\pi^{OIM}}(x_1, a_1) - 2\epsilon_1$$
$$\geq Q_{\tilde{M}}^{\pi^*}(x_1, a_1) - 3\epsilon_1 \geq Q^*(x_1, a_1) - 4\epsilon_1$$
$$= Q^*(x_1, a_1) - \epsilon,$$

where we applied (in this order) the property that truncation decreases the value function; Eq. (4); our assumption; the ϵ_1-horizon property of H; Eq. (3); Lemma 4.2 and the definition of ϵ_1.

5. Experiments

To assess the practical utility of OIM, we compared its performance to other exploration methods. Experiments were run on several small benchmark tasks challenging exploration algorithms.

For fair comparisons, benchmark problems were taken from the literature without changes, nor did we change the experimental settings or the presentation of experimental data. It also means that the presentation format varies for different benchmarks.

5.1. RiverSwim and SixArms

The first two benchmark problems, *RiverSwim* and *SixArms*, were taken from Strehl and Littman (2006).

The *RiverSwim* MDP has 6 states, representing the position of the agent in a river. The agent has two

Table 1. Results on the *RiverSwim* task.

Method	Cumulative reward
E^3	$3.020{\cdot}10^6 \pm 0.027 \cdot 10^6$
R-MAX	$3.014{\cdot}10^6 \pm 0.039 \cdot 10^6$
MBIE(SL)	$3.168{\cdot}10^6 \pm 0.023 \cdot 10^6$
MBIE-EB	$3.093{\cdot}10^6 \pm 0.023 \cdot 10^6$
OIM	$3.201{\cdot}10^6 \pm 0.016 \cdot 10^6$

Table 2. Results on the *SixArms* task.

Method	Cumulative reward
E^3	$1.623{\cdot}10^6 \pm 0.244 \cdot 10^6$
R-MAX	$2.819{\cdot}10^6 \pm 0.256 \cdot 10^6$
MBIE(SL)	$9.205{\cdot}10^6 \pm 0.559 \cdot 10^6$
MBIE-EB	$9.486{\cdot}10^6 \pm 0.587 \cdot 10^6$
OIM	$10.007{\cdot}10^6 \pm 0.654 \cdot 10^6$

possible actions: she can swim either upstream or downstream. Swimming down is always successful, but swimming up succeeds only with a 30% chance and there is a 10% chance of slipping down. The lowermost position yields +5 reward per step, while the uppermost position yields +10000.

The *SixArms* MDP consists of a central state and six 'payoff states'. In the central state, the agent can play 6 one-armed bandits. If she pulls arm k and wins, she is transferred to payoff state k. Here, she can get a reward in each step, if she chooses the appropriate action. The winning probabilities range from 1 to 0.01, while the rewards range from 50 to 6000 (for the exact values, see Strehl & Littman, 2006).

Data for E^3, R-MAX, MBIE and MBIE-EB are taken from Strehl and Littman (2006). Parameters of all four algorithms were chosen optimally. Following a coarse search in parameter space, the $R_{\max}$ parameter for OIM was set to 2000 for *RiverSwim* and to 10000 for *SixArms*. State spaces are small and value iteration instead of prioritized sweeping was completed in each step.

On both problems, each algorithm ran for 5000 time steps and the undiscounted total reward was recorded. The averages and 95% confidence intervals are calculated over 1000 test runs (Tables 5.1 and 5.1).

5.2. 50×50 Maze with Subgoals

Another benchmark problem, *MazeWithSubgoals*, was suggested by Wiering and Schmidhuber (1998). The agent has to navigate in a 50×50 maze from the start position at $(2, 2)$ to the goal (with +1000 reward) at the opposite corner $(49, 49)$. There are sub-

Table 3. Results on the *MazeWithSubgoals* task. The number of steps required to learn *p*-optimal policies (*p*=0.95, 0.99, 0.998) on the 50×50 maze task with suboptimal goals. In parentheses: how many runs out of 20 have found the goal. 'k' stands for 1000.

Method	95%	99%	99.8%
ε-GREEDY, ε = 0.2	– (0)	– (0)	– (0)
ε-GREEDY, ε = 0.4	43k (4)	52k (4)	68k (4)
RECENCY-BONUS	27k (19)	55k (18)	69k (9)
FREQ.-BONUS	24k (20)	50k (16)	66k (10)
MBIE(WS)	25k (20)	42k (19)	66k (18)
OIM	19k (20)	29k (20)	31k (20)

Table 4. Average accumulated rewards on the *Chain* task. Optimal policy gathers 3677.

METHOD	PHASE 1	PHASE 2	PHASE 8
QL+VAR.-BONUS	–	2570[1]	–
QL+ERR.-BONUS	–	2530[1]	–
QL ε-GREEDY	1519	1611	1602
QL BOLTZMANN	1606	1623	–
IEQL+	2344	2557	–
BAYESIAN QL	1697	2417	–
BAYESIAN DP[2]	3158	3611	3643
OIM	3510	3628	3643

optimal goals (with +500 reward) at the other two corners. The maze has blocked places and punishing states (−10 reward), set randomly in 20-20% of the squares. The agent can move in four directions, but with a 10% chance, its action is replaced by a random one. If the agent tries to move to a blocked state, it gets a reward of −2. Reaching any of the goals resets the agent to the start state. In all other cases, the agent gets a −1 reward for each step.

Each algorithm was run on 20 different mazes for 100,000 steps. After every 1000 steps, we tested the learned value functions by averaging 20 test runs, in each one following the greedy policy for 10,000 steps, and averaging cumulated (undiscounted) rewards. We measured the number of test runs needed for the algorithms to learn to collect 95%, 99% and 99.8% of the maximum possible rewards in 100,000 steps, and the number of steps this takes on average, if the algorithms can meet the challenge.

The algorithms that we compared were the recency based and frequency based exploration bonus methods, two versions of ε-greedy exploration, MBIE(WS) and OIM. All exploration rules applied the improved prioritized sweeping of Wiering and Schmidhuber (1998). OIM's R_{max} was set to 1000. The results are summarized in Table 3.

5.3. Chain, Loop and FlagMaze

The next three benchmark MDPs, the *Chain*, *Loop* and *FlagMaze* tasks were investigated, e.g., by Meuleau and Bourgine (1999), Strens (2000) and Dearden (2000). In the *Chain* task, 5 states are lined up along a chain. The agent gets +2 reward for being in state 1 and +10 for being in state 5. One of the actions advances one state ahead, the other one resets the agent to state 1. The *Loop* task has 9 states in

two loops (arranged in a 8-shape). Completing the first loop (using any combination of the two actions) yields +1 reward, while the second loop yields +2, but one of the actions resets the agent to the start. The *FlagMaze* task consists of a 6 × 7 maze with several walls, a start state, a goal state and 3 flags. Whenever the agent reaches the goal, her reward is the number of flags collected.

The following algorithms were compared: Q-learning with variance-based and TD error-based exploration bonus (model-free variants), ε-greedy exploration, Boltzmann exploration, IEQL+, Bayesian Q-learning, Bayesian DP and OIM. Data were taken from Meuleau and Bourgine (1999), Strens (2000) and Dearden (2000). According to the sources, parameters for all algorithms were set optimally. OIM's R_{max} parameter was set to 0.5, 10 and 0.005 for the three tasks, respectively.

Each algorithm ran for 8 learning phases. The total cumulated reward over each learning phase was measured. One phase lasted for 1000 steps for the first two tasks and 20,000 steps for the *FlagMaze* task. We carried out 256 parallel runs for the first 2 tasks and 20 for the third one.

6. Summary of the Results

We proposed a new algorithm for exploration and reinforcement learning in Markov decision processes. The algorithm integrates concepts from other advanced exploration methods. The key component of our algorithm is an optimistic initial model. The optimal policy according to the agent's model will either explore new information that helps to make the model

---

[1]Results for Phase 5.

[2]Augmented with limited amount of pre-wired knowledge (the list of successor states).

Table 5. Average accumulated rewards on the *Loop* task. Optimal policy gathers 400.

METHOD	PHASE 1	PHASE 2	PHASE 8
QL+VAR.-BONUS	–	179[1]	–
QL+ERR.-BONUS	–	179[1]	–
QL ϵ-GREEDY	337	392	399
QL BOLTZMANN	186	200	–
IEQL+	264	293	–
BAYESIAN QL	326	340	–
BAYESIAN DP[2]	377	397	399
OIM	393	400	400

Table 6. Average accumulated rewards on the *FlagMaze* task. Optimal policy gathers approximately 1890.

METHOD	PHASE 1	PHASE 2	PHASE 8
QL ϵ-GREEDY	655	1135	1147
QL BOLTZMANN	195	1024	–
IEQL+	269	253	–
BAYESIAN QL	818	1100	–
BAYESIAN DP[2]	750	1763	1864
OIM	1133	1169	1171

more accurate, or follows a near-optimal path. The extent of optimism regulates the amount of exploration. We have shown that with a suitably optimistic initialization, our algorithm finds a near-optimal policy in polynomial time. Experiments were conducted on a number of benchmark MDPs. According to the experimental results our novel method is robust and compares favorably to other methods.

Acknowledgments

References

Auer, P., & Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning algorithms. *NIPS* (pp. 49–56).

Brafman, R. I., & Tennenholtz, M. (2001). R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *Proc. IJCAI* (pp. 953–958).

Dearden, R. W. (2000). *Learning and planning in structured worlds*. Doctoral dissertation, University of British Columbia.

Even-Dar, E., & Mansour, Y. (2001). Convergence of optimistic and incremental Q-learning. *NIPS* (pp. 1499–1506).

Kearns, M., & Singh, S. (1998). Near-optimal reinforcement learning in polynomial time. *Proc. ICML* (pp. 260–268).

Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, *49*, 209–232.

Koenig, S., & Simmons, R. G. (1993). Complexity analysis of real-time reinforcement learning. *Proc. AAAI* (pp. 99–105).

Littman, M. L., & Szepesvári, C. (1996). A generalized reinforcement-learning model: Convergence and applications. *Proc. ICML* (pp. 310–318). Morgan Kaufmann.

Meuleau, N., & Bourgine, P. (1999). Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, *35*, 117–154.

Singh, S. P., Jaakkola, T., Littman, M. L., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, *38*, 287–308.

Strehl, A. L., & Littman, M. L. (2005). A theoretical analysis of model-based interval estimation. *Proc. ICML* (pp. 856–863).

Strehl, A. L., & Littman, M. L. (2006). An analysis of model-based interval estimation for Markov decision processes. *Submitted*.

Strens, M. (2000). A Bayesian framework for reinforcement learning. *Proc. ICML* (pp. 943–950). Morgan Kaufmann, San Francisco, CA.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge.

Szita, I., & Lőrincz, A. (2008). *The many faces of optimism – extended version* (Technical Report). Eötvös Loránd University, Hungary.

Wiering, M. A., & Schmidhuber, J. (1998). Efficient model-based exploration. *Proc. SAB: From Animals to Animats* (pp. 223–228).