Optimism in the Face of Uncertainty Should be Refutable

Ronald ORTNER

Montanuniversität Leoben
Department Mathematik und Informationstechnolgie
Franz-Josef-Strasse 18, 8700 Leoben, Austria,
Phone number: ++43 (0)3842 402-1503,
fax: ++43 (0)3842 402-1502,
e-mail: rortner@unileoben.ac.at

September 17, 2008

Abstract

We give an example from the theory of Markov decision processes which shows that the "optimism in the face of uncertainty" heuristics may fail to make any progress. This is due to the impossibility to falsify a belief that a (transition) probability is larger than 0. Our example shows the utility of Popper's demand of falsifiability of hypotheses in the area of artificial intelligence.

Keywords: Markov decision processes, refutability, reinforcement learning.

1 Introduction

In a Markov decision process (short MDP), an agent operates on a set of distinguishable states S. In each state the agent has a number of actions taken from a set A at her disposal. The set of available actions A may depend on the current state, but usually A is chosen to be the set of actions available in each state. According to the chosen action the agent obtains some reward. Typically this is a real number taken from some interval $[0, R]$. Commonly, the rewards are not deterministic but random according to probability distributions which depend on the current state and the chosen action. Further, after having chosen an action a in state s, the agent is transferred to another state according to a probability distribution depending on state s and action a. The agent starts from a fixed state (or, more generally, is randomly set to some initial state according to a given probability distribution over the set of states).

A formal definition of a Markov decision process thus may look as follows.

Definition 1 *A* Markov decision process *(MDP)* $\mathcal{M}$ *on a finite set of* states S *with a finite set of* actions A *available in each state in* S *consists of*

(i) an initial state s_0*, or more generally an initial distribution* μ_0 *over* S*,*

(ii) the transition probabilities $p(s'|s, a)$ *that specify the probability of reaching state* s' *when choosing action* a *in state* s*, and*

(iii) the payoff distributions with mean $r(s, a)$ *and support in* $[0, R]$ *that specify the random reward for choosing action* a *in state* s*.*

A *(stationary) policy* on an MDP $\mathcal{M}$ is a mapping $\pi : S \to A$, which for each state specifies the action the agent chooses. For a fixed policy the random process induced on the states (which is a *Markov chain*) may exhibit different behavior. In the simplest case the induced Markov chain is *ergodic*, that is, with positive probability each state will be visited sooner or later (with probability 1), independent of the initial state. However, there may be also *transient* states, which independent of the initial state will (with probability 1) be visited only a finite number of times. Finally, the set of non-transient states may be partitioned into several *communicating classes* such that there are no positive transition probabilities between states in different classes. In any case, the *average reward of a policy* π can be defined as

$$\rho(\mathcal{M},\pi) \;:=\; \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big(r\big(s^{(t)},\pi(s^{(t)})\big)\Big),$$

where $s^{(t)}$ is the (random) state that is visited at step t. One can show that the average reward of any policy π can be written as

$$\rho(\mathcal{M},\pi) \;=\; \sum_{s\in S}\mu_\pi(s)\,r\big(s,\pi(s)\big),$$

where μ_π is the stationary distribution of the Markov chain which π induces on $\mathcal{M}$. Intuitively, the stationary distribution on S indicates the probability of being in a particular state after an infinite number of steps. More precisely, for an ergodic Markov chain with transition matrix[1] $P = \big(p(s,s')\big)_{s,s'\in S}$ there exists a unique invariant and strictly positive distribution μ_π, such that independent of μ_0 it holds that $\mu_0\bar{P}_n$ converges to μ_π for $n\to\infty$, where

[1] The transition matrix P is the matrix of transition probabilities with rows and columns indexed by the states in S, so that in row s and column s' the entry in P is the transition probability $p(s,s')$ from s to s'.

$\bar{P}_n = \frac{1}{n}\sum_{j=1}^{n} P^j$. If the Markov chain is not ergodic, μ_π will depend on the initial state s_0, or more generally, the initial distribution μ_0 [Kemeny et al., 1976].

The goal of an agent operating on an MDP is to maximize her average reward, that is, to find an *optimal policy* π^* such that $\rho(\mathcal{M},\pi^*) \geq \rho(\mathcal{M},\pi)$ for all policies π. It can be shown that the achievable average reward cannot be increased by using time-dependent policies [Puterman, 1994].

2 The "Optimism in the Face of Uncertainty" Maxim

If the MDP is known to the learner, there are various algorithms to calculate an optimal policy, such as value iteration or policy iteration [Puterman, 1994]. The problem becomes more interesting if one assumes that the agent — beside her knowledge of S and A — can only observe the current state and the rewards obtained for an action. In this setting, many algorithms follow the "optimism in the face of uncertainty" maxim, which lets the agent act according to an overly optimistic model of the MDP with respect to the observations so far.

The best example for such an algorithm is the well-known R-Max algorithm [Brafman and Tennenholtz, 2002]. The agent's model of the MDP assumes the maximal possible reward R for each action a in states in which a has not been probed sufficiently often. Here "sufficiently often" means that if all actions in all states have been visited sufficiently often, then the optimal policy with respect to the estimated MDP (i.e. the MDP whose rewards and transition probabilities are the means of the obtained rewards and observed transitions, respectively) will be (close to) optimal in the real MDP as well. The idea of the algorithm is that the agent will explore states and actions that are not known well enough (at least if it may pay off to do so). That way the optimistic model is an incentive to explore.

Two similar algorithms which refine the idea R-Max is based on are MBIE [Strehl and Littman, 2004, 2005] and UCRL [Auer and Ortner, 2006]. Here the agent assumes the most optimistic model with respect to some confidence intervals on the estimated transition probabilities and rewards.

3 An Example where Optimism Fails

3.1 The Example

The following example shows that the "optimism in the face of uncertainty" maxim may fail to find an optimal policy.

Example 1 *Assume an MDP with two states s_1 and s_2, where s_1 is the initial state. In s_1 two actions a_1 and a_2 are available, none of which however results in a transition to s_2 (cf. Figure 1).*

Even if the two actions yield different average reward $< R$, any optimistic algorithm will choose a model in which there is a positive transition probability

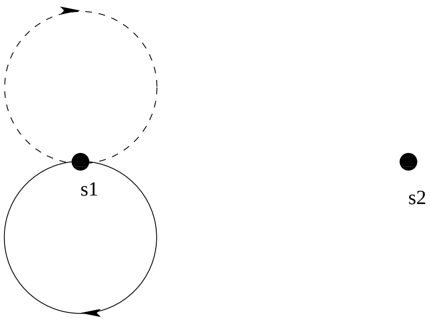

Figure 1: The MDP of Example 1.

to s_2, where the algorithm optimistically expects to yield maximal reward and the possibility to stay, i.e. $r(s_2, a) = R$ and $p(s_2|s_2, a) = 1$ for some action a. With this model, any algorithm will choose not the (optimal) action that gives highest reward in state s_1, but rather the action it expects to cause a transition to s_2. Note that it is not important how small the transition probability from s_1 to s_2 is estimated. Even the smallest positive probability results in a stationary distribution μ with $\mu(s_1) = 0$ and $\mu(s_2) = 1$, so that in the long run it does not matter how much time is used to insist in the transition to s_2.

The problem for the algorithm is that it is impossible to distinguish between a very low probability for a transition and its impossibility. Here the "optimism in the face of uncertainty" idea fails, as there is no way to falsify the wrong belief in a possible transition. [2] [3]

[2] Note that the claim that the outcome of a random experiment has positive probability is basically an existence claim over an infinite number of trials (and hence not refutable).

[3] Incidentally, this can also be used as criticism on *Pascal's wager*. Although it is not clear whether it is appropriate to represent Pascal's wager as an MDP similar to that in Example 1, Pascal's argument is based on a non-refutable belief, as the assumption that

This example is reminiscent of well-known cases where chess programs fail to evaluate totally blocked positions correctly (cf. Figure 2). While any human chess player immediately will recognize these kind of positions to be deadly drawn (even independent of what any of the players may do in the future), programs, which usually evaluate positions according to the material distribution, do not see that there is no possibility to exploit the material advantage.

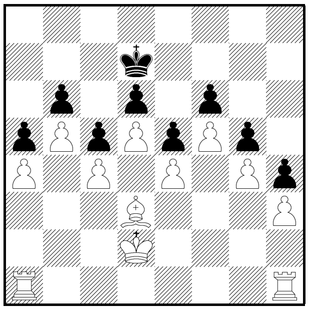

Figure 2: A chess position which is drawn, as there is no possibility for White to exploit its material advantage. Note that White cannot even sacrifice material in order to break up the pawn chain.

3.2 How to Avoid the Problem

How do the optimistic algorithms mentioned above deal with this problem? First, most algorithms such as MBIE [Strehl and Littman, 2004, 2005] do not work with average but with discounted rewards, that is, at step t the expected reward the agent receives for action a is not the full amount of $r(s,a)$ but only $\gamma^t \cdot r(s,a)$ for some discount factor $\gamma < 1$. This means that

there is a positive transition probability to heaven is not falsifiable.

basically only a finite number of steps matter for the agent. However, in that case one may calculate for how many steps it pays off to insist in a transition which has never been observed.[4]

The algorithm R-Max [Brafman and Tennenholtz, 2002] avoids the problem of Example 1 basically by considering only states which have been visited sufficiently often. Further, knowledge of an MDP dependent parameter, the *ε-return mixing time*, is assumed. This is the time it takes until the agent's reward is ε-close to the optimal average reward, when playing an optimal policy. Obviously, this parameter also indicates how small transition probabilities under an optimal policy might be.

Similarly to R-Max one may try to ignore the knowledge of the whole state space and consider only states which have been visited before. However, this meets the difficulty of transient states as the following example shows.

Example 2 *Consider the same MDP as in Example 1, only that now the initial state is s_2, in which as in s_1 the agent has two actions a_1, a_2 at her disposal, both of which lead with probability 1 to s_1 (cf. Figure 3).*

The agent chooses e.g. a_1, and gets an arbitrary reward. However, the optimistic assumptions about action a_2 are the same as in example 1, and we have the same problem.

[4] That is, for an optimistic estimate of the transition probability p in question, one computes the expected reward which may be gained when insisting in the transition. This can be compared to the reward to be expected when ignoring the transition (i.e. setting $p = 0$ in the agent's model). If the latter value is larger, the agent refutes the hypothesis that $p > 0$.

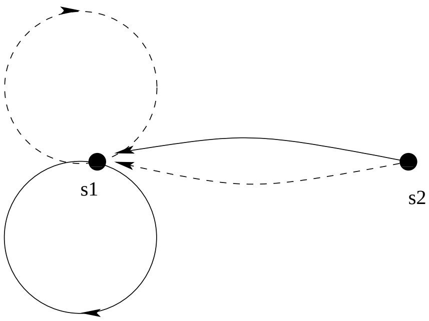

Figure 3: The MDP of Example 2.

Note that the MDPs in the examples are not ergodic. Actually, the problem of Examples 1 and 2 cannot occur in ergodic MDPs. Thus, e.g. the UCRL algorithm [Auer and Ortner, 2006] only considers ergodic MDPs. Recently, the results of [Auer and Ortner, 2006] have been generalized to *communicating* MDPs in which any two states can be reached from each other within a finite number of steps under some policy [Auer et al., 2008]. However, the MDPs of Examples 1 and 2 are not communicating either.[5]

3.3 Popper, Refutability, and the Progress of Science

So far, our notion of refutation has been only intuitive. Usually, refutation means that the observations contradict the hypothesis (in a strictly logical sense). However, as we are dealing with probabilities, in our context refutation rather means that a current hypothesis is quite improbable given the observations. That is, there is a threshold for the probability of the hypoth-

[5] It is worth noting that although the UCRL algorithm assumes that the underlying MDP is ergodic or communicating, the optimistic model of the MDP it assumes in general is neither ergodic nor communicating.

esis given the observations, which decides whether a hypothesis is kept or refuted. This is basically also what Popper had in mind for statistical hypotheses (cf. [Popper, 1969], Chap. VIII). Of course, in the MDP framework there is no global threshold which will work for any MDP, as transition probabilities may be arbitrarily small, so that this approach cannot give a general solution for the problems posed by Examples 1 and 2.

Still, Popper's notion of refutability at least provides some explanation why the "optimism in the face of uncertainty" heuristics fails in these examples. For Popper refutability was on the one hand a criterion for empirical hypotheses which separates them from metaphysical hypotheses (cf. [Popper, 1969], Chap. IV). On the other hand, he thought that refutability of hypotheses would be a warranty for scientific progress towards truth (cf. [Popper, 1969], Chap. X). We think that our examples show that Popper's theory holds in the nutshell of Markov decision processes, even if the agent operating on the MDP wants to converge to *optimal* rather than *absolute* truth.

What makes refutable optimism work is that in the case where the agent's hypothesis is not refuted she is doing rather well. That's why refutable optimism is preferred to other refutable hypotheses. Unfortunately, as Examples 1 and 2 show, in some cases optimism becomes irrefutable, while giving up optimism may always happen one step too early.

Acknowledgments.

The author would like to thank Georg Dorn for comments on a prior version of this paper. This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Auer, P. & Ortner, R. (2006). Logarithmic online regret bounds for reinforcement learning. In B. Schölkopf, J.C. Platt, & T. Hofmann (Eds.), Advances in Neural Information Processing Systems 19 (pp. 49–56). Cambridge, MA: MIT Press.

Auer, P., Jaksch, T., & Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. submitted.

Brafman, R. I. & Tennenholtz, M. (2002) R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3, 213–231.

Kemeny, J.G., Snell, J.L., & Knapp, A.W. (1976) *Denumerable Markov Chains.* New York: Springer.

Popper, K. R. (1969). *Logik der Forschung.* Tübingen: Mohr, third edition.

Puterman, M. L (1994). *Markov Decision Processes. Discrete Stochastic Programming.* New York: Wiley.

Strehl, A. L. & Littman, M. L. (2004). An empirical evaluation of interval estimation for Markov decision processes. In 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004) (pp. 128–135). IEEE Computer Society.

Strehl, A. L. & Littman, M. L. (2005). A theoretical analysis of model-based interval estimation. In L. De Raedt, S. Wrobel (Eds.), Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005) (pp. 857–864). ACM.