Adaptive control without prior by dynamic programming

Matthias C. M. Troffaes[1]
SYSTeMS Research Group, Universiteit Gent
Technologiepark Zwijnaarde 914, 9052 Zwijnaarde, Belgium
Matthias.Troffaes@UGent.be

1 Introduction

Adaptive control is often subject to serious bias in the learning phase, simply because insufficient information is available in order to motivate the choice of a unique prior, and hence, a unique optimal feedback. Imprecise probability theory may resolve this problem by means of using sets of priors. In doing so, we end up with a set of possibly optimal feedback controls—rather than a single one. This allows us to quantify the lack of information for finding an optimal feedback, and tells us exactly how many transitions must be observed before we can have a unique, robust optimal feedback. Such a result would be especially useful in applications where sampling costs are relatively high compared to the rewards incurred at each transition.

Adaptive control of Markov decision processes with uncertain transition probabilities has already been studied in great detail during the sixties [2]. In the classical approach to this control problem, the uncertainty of the transition probabilities is described by means of a product of Dirichlet priors, which are updated in time as transitions are observed. It is well-known that the optimal solution can be found through a dynamic programming algorithm.

Renewed interest in this problem has been initiated by recent developments in imprecise probability theory. It has been demonstrated how we can learn about the probabilities of a multinomial sampling model without having to give a unique prior, by means of a set of Dirichlet priors [3]. In optimal control, it turns out that the dynamic programming formalism still applies to dynamical systems whose gain is described by a set of probability distributions [1]. These results are our main inspiration for generalising adaptive control of Markov decision processes with uncertain transition probabilities to the framework of imprecise probabilities.

2 Example

Consider the Markov decision process depicted in Figure 1. At each time k we can choose between two actions, u and v. Transition probabilities are denoted as p^v_{yx} (the probability from state y to state x when taking action v), and the reward

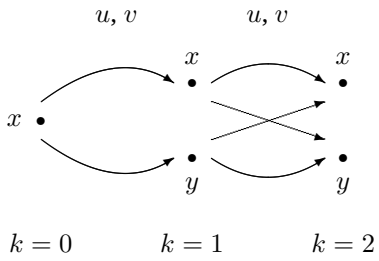

Figure 1: A simple Markov decision process

associated with this transition is denoted by r^v_{yx}. Transition probabilities are unknown, we only know the rewards, *e.g.*

$$r^u_{xx} = r^u_{yx} = 1 \qquad r^v_{xx} = r^v_{yx} = 2$$
$$r^u_{xy} = r^u_{yy} = 1.5 \qquad r^v_{xy} = r^v_{yy} = 0.75$$

Intuitively, it is clear that insufficient information is available in order to construct a unique optimal feedback. Suppose we are in state x at time $k = 0$, take action v and end up in state x at time $k = 1$. Then it seems reasonable to assume that when we select action v again, the probability that we end up in x again is higher than the probability of ending up in y. In fact, the reward associated with this transition, r^v_{xx}, is the highest possible reward. Even if we do not know precisely the value of p^v_{xx}, after observing the transition from state x at time k to state x at $k + 1$ under action v, we obtain, through the imprecise Dirichlet model (hyperparameter $s = 1$), a sufficiently narrow probability interval for p^v_{xx} in order to ensure that we will end up with the highest possible reward by taking action v from state x at time $k = 1$. Secondly, we have found that this model satisfies the principle of optimality. So, globally optimal feedback controls can be obtained through an efficient dynamic programming algorithm.

[1] This paper presents research results of project G.0139.01 of the Fund for Scientific Research, Flanders (Belgium), and of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian state, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with the author.

References

[1] Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for discrete-time systems with uncertain gain. In Jean-Marc Bernard et al., editors, *ISIPTA '03 – Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, pages 162–176. Carleton Scientific, July 2003.

[2] J. J. Martin. *Bayesian Decision Theory and Markov Chains*. John Wiley & Sons, New York, 1967.

[3] Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society*, 58(1):3–34, 1996.