Nearly Optimal Exploration-Exploitation Decision Thresholds

Christos Dimitrakakis *

IDIAP Research Institute, 4 Rue de Simplon, Martigny CH 1920, Switzerland
dimitrak@idiap.ch

Abstract. While in general trading off exploration and exploitation in reinforcement learning is hard, under some formulations relatively simple solutions exist. Optimal decision thresholds for the multi-armed bandit problem, one for the infinite horizon discounted reward case and one for the finite horizon undiscounted reward case are derived, which make the link between the reward horizon, uncertainty and the need for exploration explicit. From this result follow two practical approximate algorithms, which are illustrated experimentally.

1 Introduction

In reinforcement learning, the dilemma between selecting actions to maximise the expected return according to the current world model and to improve the world model such as to *potentially* be able to achieve a higher expected return is referred to as the *exploration-exploitation trade-off*. This has been the subject of much interest before, one of the earliest developments being the theory of sequential sampling in statistics, as developed by [1]. This dealt mostly with making sequential decisions for accepting one among a set of particular hypotheses, with a view towards applying it to jointly decide the termination of an experiment and the acceptance of a hypothesis. A more general overview of sequential decision problems from a Bayesian viewpoint is offered in [2].

The optimal, but intractable, Bayesian solution for bandit problems was given in [3], while recently tight bounds on the sample complexity of exploration have been found [4]. An approximation to the full Bayesian case for the general reinforcement learning problem is given in [5], while an alternative technique based on eliminating actions which are confidently estimated as low-value is given in [6].

The following section formulates the intuitive concept of trading exploration and exploitation as a natural consequence of the *definition* of the problem of reinforcement learning. After the problem definitions which correspond to either extreme are identified, Sec. 3 derives a threshold for switching from exploratory to greedy behaviour in bandit problems. This threshold is found to depend on the

* Thanks to M. Keller and R. Chavarriaga, for comments and interesting discussions. This work has received financial support from the Swiss NSF under the MULTI project (2000-068231.021/1) and from IDIAP.

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 850–859, 2006.

effective reward horizon of the optimal policy and on our current belief distribution of the expected rewards of each action. A sketch of the extension to MDPs is presented in Sec. 4. Section 5 uses an upper bound on the value of exploration to derive practical algorithms, which are then illustrated experimentally in Sec. 6. We conclude with a discussion on the relations with other methods.

2 Exploration Versus Exploitation

Let us assume a standard multi-armed bandit setting, where a reward distribution $p(r_{t+1}|a_t)$ is conditioned on actions in $a_t \in \mathcal{A}$, with $r_t \in \mathbb{R}$. The aim is to discover a policy $\pi = \{P(a_t = i)|i \in \mathcal{A}\}$, where $P(a_t = i)$ is the probability that action i is chosen at time t, which maximises $E[r_{t+1}|\pi]$, the expected value of the reward at the following time-step under the distribution defined by the policy π. It follows that the optimal gambler, or oracle, for this problem would be a policy which always chooses $i \in \mathcal{A}$ such that $E[r_{t+1}|a_t = i] \geq E[r_{t+1}|a_t = j]$ for all $j \in \mathcal{A}$. Given the conditional expectations, implementing the oracle is trivial. However this tells us little about the optimal way to select actions when the expectations are unknown. As it turns out, the optimal action selection mechanism will depend upon the problem formulation. We initially consider the two simplest cases in order to illustrate that the exploration/exploitation tradeoff is and should be viewed in terms of problem and model definition.

In the first problem formulation the objective is to discover a parameterized probabilistic policy $\pi = \big\{P(a_t|\theta_t) \bigm| a_t \in \mathcal{A}\big\}$, with parameters θ_t, for selecting actions such that $E[r_{t+1}|\pi]$ is maximised. If we consider a model whose parameters are the set of estimates $\theta_t = \big\{q_i = \hat{E}_t[r_{t+1}|a_t = i] \bigm| i \in \mathcal{A}\big\}$, then the optimal choice is to select a_t for which the estimated expected value of the reward is highest, because according to our current belief any other choice will necessarily lead to a lower expectation. Thus, stating the bandit problem in this way does not allow the exploration of seemingly lower, but potentially higher value actions and it results in a *greedy* policy.

In the second formulation, we wish to minimise the discrepancy between our estimate q_i and the true expectation. This could be written as the following minimisation problem:

$$\sum_{i \in \mathcal{A}} E\big[\|r_{t+1} - q_i\|^2 \bigm| a_t = i\big].$$

For point estimates of the expected reward, this requires sampling *uniformly* from all actions and thus represents a purely exploratory policy. If the problem is stated as simply minimising the discrepancy asymptotically, then uniformity is not required and it is only necessary to sample from all actions infinitely often. This condition holds when $P(a_t = i) > 0 \ \forall i \in \mathcal{A}, \ t > 0$ and can be satisfied by mixing the optimal policies for the two formulations, with a probability ϵ of using the uniform action selection and a probability $1 - \epsilon$ of using the greedy action selection. This results in the well-known ϵ-greedy policy (see for example [7]), with the parameter $\epsilon \in [0, 1]$ used to control exploration.