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.

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.

This formulation of the exploration-exploitation problem, though leading to an intuitive result, does not lead to an obvious way to optimally select actions. In the following section we shall consider bandit problems for which the functional to be maximised is

$$E\left[\sum_{k=0}^{N} g(k) r_{t+k+1} \middle| \pi\right], \quad g(k) \in [0,1],\ N \geq 0,$$

with $\sum_{k=0}^{\infty} g(k) < \infty$. In this formulation of the problem we are not only interested in maximising the expected reward at the next time step, but in the subsequent N steps, with the $g(\cdot)$ function providing another convenient way to weigh our preference among short and long-term rewards. Intuitively it is expected that the optimal policy for this problem will be different depending on how long-term are the rewards that we are interested in. As will be shown later, by lengthening the effective reward horizon through manipulation of g and N, i.e. by changing the definition of the problem that we wish to solve, the exploration bias is increased automatically.

3 Optimal Exploration Threshold for Bandit Problems

We want to know when it is a better decision to take action i rather than some other action j, with $i, j \in \mathcal{A}$, given that we have estimates q_i, q_j for $E[r_{t+1}|a_t = i]$ and $E[r_{t+1}|a_t = j]$ respectively[1]. We shall attempt to see under which conditions it is better to take an action different than the one whose expected reward is greatest. For this we shall need the following assumption:

Assumption 1 (Expected rewards are bounded from below). *There exists $b \in \mathbb{R}$ such that*

$$E[r_{t+1}|a_t = i] \geq b \quad \forall\ i \in \mathcal{A}, \tag{1}$$

The above assumption is necessary for imposing a lower bound on the expected return of exploratory actions: no matter what action is taken, we are guaranteed that $E[r_t] > b$. Without this condition, exploratory actions would be too risky to be taken at all.

Given two possible actions to take, where one action is currently estimated to have a lower expected reward than the other, then it might be worthwhile to pursue the lower-valued action if the following conditions are true: (a) there is a degree of uncertainty such that the lower-valued action can potentially be better than the higher-valued one, (b) we are interested in maximising more than just the expectation of the next reward, but the expectation of a weighted sum of future rewards, (c) we will be able to accurately determine whether one action is better than the other quickly enough, so that not a lot of resources will be wasted in exploration.

[1] For bandit problems with states in a state space $\mathcal{S}$, similar arguments can be made by considering $i, j \in \mathcal{S} \times \mathcal{A}$.

We now start viewing q_i as random variables for which we hold belief distributions $p(q_i)$, with $\bar{q}_i = E[q_i] = \hat{E}[r_{t+1}|a_t = i]$. The problem can be defined as deciding when action i, is better than taking action j, under the condition that doing so allows us to determine whether $q_i > q_j + \delta$ with high probability after $T \geq 1$ exploratory actions. For this reason we will need the following bound on the expected return of exploration.

Lemma 1 (Exploration bound). *For any return of the form $R_t = \sum_{k=0}^{N} g(k) r_{t+k+1}$, with $g(k) \geq 0$, assuming (1) holds, the expected return of taking action i for T time-steps and following a greedy policy thereafter, when $\bar{q}_i > \bar{q}_j$, is bounded below by*

$$U(i,j,T,\delta,b) = \sum_{k=T}^{N} g(k)\big((\bar{q}_j+\delta)P(q_i > q_j+\delta) + \bar{q}_j P(q_i \leq q_j+\delta)\big) + \sum_{k=0}^{T-1} g(k)\big((\bar{q}_j+\delta)P(q_i > q_j+\delta) + bP(q_i \leq q_j+\delta)\big) \tag{2}$$

for some $\delta > 0$.

This follows immediately from Assumption 1. The greedy behaviour supposes we are following a policy where we continue to perform i if we know that $P(q_i > q_j + \delta) \approx 1$ after T steps and switch back to j otherwise.

Without loss of generality, in the sequel we will assume that $b = 0$ (If expected rewards are bounded by some $b \neq 0$, we can always subtract b from all rewards and obtain the same). For further convenience, we set $p_i = P(q_i \geq q_j + \delta)$. Then we may write that we must take action i if the expected return of simply taking action j is smaller than the expected return of taking action i for T steps and then behaving greedily, i.e. if the following holds:

$$\sum_{k=0}^{N} g(k)\bar{q}_j < \sum_{k=T}^{N} g(k)\big((\bar{q}_j+\delta)p_i + \bar{q}_j(1-p_i)\big) + \sum_{k=0}^{T-1} g(k)(\bar{q}_j+\delta)p_i \tag{3}$$

$$\sum_{k=0}^{T-1} g(k)\big(\bar{q}_j - (\bar{q}_j+\delta)p_i\big) < \sum_{k=T}^{N} g(k)\big(\delta p_i\big) \tag{4}$$

Let $g(k) = \gamma^k$, with $\gamma \in [0,1]$. In this case, any choice of T can be made equivalent to $T = 1$ by dividing everything with $\sum_{k=0}^{T-1}\gamma^k$. We explore two cases: $\gamma < 1,\ N \to \infty$ and $\gamma = 1,\ N < \infty$. In the first case, which corresponds to infinite horizon exponentially discounted reward maximisation problems, we obtain the following:

$$\bar{q}_j - (\bar{q}_j+\delta)p_i < \sum_{k=1}^{\infty} \gamma^k \delta p_i \tag{5}$$

$$\frac{\bar{q}_j - (\bar{q}_j+\delta)p_i}{(1-p_i)\bar{q}_j} < \gamma. \tag{6}$$

It is possible to simplify this expression considerably. When $P(q_i \geq \bar{q_j}+\delta) = 1/2$, it follows from (6) that

$$\gamma > \frac{\bar{q_j} - (\bar{q_j} + \delta)/2}{\bar{q_j}/2} = \frac{\bar{q_j} - \delta}{\bar{q_j}}. \tag{7}$$

Thus, for infinite horizon discounted reward maximisation problems, when it is known that the all expected rewards are non-negative, all we need to do is find δ such that $P(q_i \geq q_j + \delta) = 1/2$. Then (7) can be used to make a decision on whether it is worthwhile to perform exploration. Although it might seem strange that q_i is omitted from this expression, its distribution is implicitly expressed through the value of δ.

In the second case, finite horizon cumulative reward maximisation problems, exploration should be performed when the following condition is satisfied:

$$N\delta p_i > \bar{q_j} - (\bar{q_j} + \delta)p_i \tag{8}$$

Here the decision making function is of a different nature, since it depends on both estimates. However, in both cases, the longer the effective horizon becomes and the larger the uncertainty is, the more the bias towards exploration is increased. We furthermore note that in the finite horizon case, the backward induction procedure can be used to make optimal decisions (see [2] Sec. 12.4).

3.1 Solutions for Specific Distributions

If we have a specific form for the distribution $P(q_i > q_j + \delta)$ it may be possible to obtain analytical solutions. To see how this can be achieved, consider that from (6), we have:

$$\begin{aligned} \gamma\bar{q_j} &> \bar{q_j} - \delta\frac{p_i}{1-p_i} \\ 0 &< \delta\frac{P(q_i > q_j + \delta)}{1 - P(q_j > q_j + \delta)} - (1-\gamma)\bar{q_j}, \end{aligned} \tag{9}$$

recalling that all mean rewards are non-negative.

If this condition is satisfied for some δ then exploration must be performed. We observe that if the first term is maximised for some δ^* for which the inequality is not satisfied, then there is no $\delta \neq \delta^*$ that can satisfy it. Thus, we can attempt to examine some distributions for which this δ^* can be determined. We shall restrict ourselves to distributions that are bounded below, due to Assumption 1.

3.2 Solutions for the Exponential Distribution

One such distribution is the exponential distribution, defined as

$$P(X > \delta) = \int_{\delta}^{\infty} \beta e^{-\beta(x-\mu)} dx = e^{-\beta(\delta-\mu)}$$

if $\delta > \mu$, 1 otherwise. We may plug this into (9) as follows

$$f(\delta) = \delta \frac{P(q_i > q_j + \delta)}{1 - P(q_i > q_j + \delta)} = \delta \frac{e^{-\beta_i(\mu_j + \delta - \mu_i)}}{1 - e^{-\beta_i(\mu_j + \delta - \mu_i)}} = \frac{\delta}{e^{\beta_i(\mu_j + \delta - \mu_i)} - 1}$$

Now we should attempt to find $\delta^* = \arg\max_\delta f(\delta)$. We begin by taking the derivative with respect to δ. Set $g(\delta) = e^{h(\delta)} - 1$, $h(\delta) = \beta_i(\bar{q}_j + \delta - \mu_i)$

$$\nabla f(\delta) = \frac{g(\delta) - \delta \nabla g(\delta)}{g(\delta)^2} = \frac{g(\delta) - \delta \beta_i \nabla_h g(\delta)}{g(\delta)^2} = \frac{e^{h(\delta)}(1 - \delta\beta_i) - 1}{(e^{h(\delta)} - 1)^2}$$

Necessary and sufficient conditions for some point δ^* to be a local maximum for a continuous differentiable function $f(\delta)$ are that $\nabla_\delta f(\delta^*) = 0$ and $\nabla^2_\delta f(\delta^*) < 0$. The necessary condition for δ results in

$$e^{\beta_i(q_k + \delta - \mu_i)}(1 - \delta\beta_i) = 1. \tag{10}$$

Unfortunately (10) has no closed form solution, but it is related to the Lambert W function for which iterative solutions do exist [8]. The found solution can then be plugged into (9) to see whether the conditions for exploration are satisfied.

4 Extension to the General Case

In the general reinforcement learning setting, the reward distribution does not only depend on the action taken but additionally on a state variable. The state transition distribution is conditioned on actions and has the Markov property. Each particular task within this framework can be summarised as a Markov decision process:

Definition 1 (Markov decision process). *A Markov decision process is defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition distribution $\mathfrak{T}(s', s, a) = P(s'_{t+1}|s_t = s, a_t = a)$ and a reward distribution $\mathfrak{R}(s', s, a) = p(r_{t+1}|s_{t+1} = s', s_t = s, a_t = a)$.*

The simplest way to extend the bandit case to the more general one of MDPs is to find conditions under which the latter reduces to the former. This can be done for example by considering choices not between simple actions but between *temporally extended actions*, which we will refer to as *options* following [9]. We shall only need a simplified version of this framework, where each possible *option* x corresponds to some policy $\pi^x : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. This is sufficient for sketching the conditions under which the equivalence arises.

In particular, we examine the case where we have two options. The first option is to always select actions according to some exploratory principle, such picking them from a uniform distribution. The second is to always select actions greedily, i.e. by picking the action with the highest expected return.

We assume that each option will last for time T. One further necessary component for this framework is the notion of mixing time

Definition 2 (Exploration mixing time). *We define the exploration mixing time for a particular MDP $\mathcal{M}$ and a policy π $T_\epsilon(\mathcal{M}, \pi)$ as the expected number of time steps after which the state distribution is close to the stationary state distribution of π after we have taken an exploratory action i at time step t, i.e. the expected number of steps T such that the following condition holds:*

$$\frac{1}{\|\mathbb{S}\|} \sum_s \|P(s_{t+T} = s | s_t, \pi) - P(s_{t+T} = s | a_t = i, s_t, \pi)\| < \epsilon$$

It is of course necessary for the MDP to be ergodic for this to be finite. If we only consider switching between options at time periods greater than $T_\epsilon(\mathcal{M}, \pi)$, then the option framework's roughly corresponds to the bandit framework, and T_ϵ in the former to T in the latter. This means that whenever we take an exploratory action i (one that does not correspond to the action that would have been selected by the greedy policy π), the distribution of states would remain to be significantly different from that under π for $T_\epsilon(\mathcal{M}, \pi)$ time steps. Thus we could consider the exploration to be taking place during all of T_ϵ, after which we would be free to continue exploration or not. Although there is no direct correspondence between the two cases, this limited equivalence could be sufficient for motivating the use of similar techniques for determining the optimal exploration exploitation threshold in full MDPs.

5 Optimistic Evaluation

In order to utilise Lemma 1 in a practical setting we must define T in some sense. The simplest solution is to set $T = 1$, which results in an optimistic estimate for exploratory actions as will be shown below. By rearranging (2) we have

$$U(i, j, T, \delta, b) = \sum_{k=0}^{N} g(k) \bar{q_j} + \sum_{k=0}^{N} g(k) \delta p_i + (1 - p_i) \left(\sum_{k=0}^{T-1} g(k)(b - \bar{q_j})) \right) \quad (11)$$

from which it is evident, since $q_j \geq b$ and $g(k) \geq 0$, that $U(i, j, T_1, \delta, b) \geq U(i, j, T_2, \delta, b)$ when $T_1 < T_2$, thus $U(i, j, 1, \delta, b) \geq U(i, j, T, \delta, b)$ for any $T \geq 1$. This can now be used to obtain Alg. 1 for optimistic exploration.

Nevertheless, testing for the existence of a suitable δ can be costly since, barring an analytic procedure it requires an exhaustive search. On the other hand, it may be possible to achieve a similar result through sampling for different values of δ. Herein, the following sampling method is considered: Firstly, we determine the action j with the greatest $\bar{q_j}$. Then, for each action i we take a sample x from the distribution $p(q_i)$ and set $\delta = x - \bar{q_j}$. This is quite an arbitrary sampling method, but we may expect to obtain a $\delta > 0$ with high probability if i has a high probability to be significantly better than j. This method is summarised in Alg. 2.

An alternative exploration method is given by Alg. 3, which samples each action with probability equal to the probability that its expected reward is the highest. It can perhaps be viewed as a crude approximation to Alg. 2 when $\gamma \to 1$ and has the advantage that it is extremely simple.

Algorithm 1 Optimistic exploration

if $\exists\, \delta \;:\; U(i, j, 1, \delta, b) > \sum_{k=0}^{N} g(k)\bar{q}_j$ **then**
  $a \Leftarrow i$
else
  $a \Leftarrow j$
end if

Algorithm 2 Optimistic stochastic exploration

$j \Leftarrow \arg\max_i \bar{q}_i$.
$u_j = \sum_{k=0}^{N} g(k)\bar{q}_j$.
for all $i \neq j$ **do**
  $\delta \Leftarrow x - \bar{q}_j, \quad x \sim p(q_i)$
  $u_i \Leftarrow U(i, j, 1, \delta, b)$
end for
$a \Leftarrow \arg\max_i u_i$

6 Experiments

A small experiment was performed on a n-armed bandit problem with rewards $r_t \in \{0, 1\}$ drawn from a Bernoulli distribution. Alg. 2 was used with $g(k) = \gamma^k$ and $b = 0$, which is in agreement with the distribution. This was compared with Alg. 3, which can be perhaps viewed as a crude approximation to Alg. 2 when $\gamma \to 1$. The performance of ϵ-greedy action selection with $\epsilon = 0.01$ was evaluated for reference.

The ϵ-greedy algorithm used point estimates for $\bar{q}_i$, which were updated with gradient descent with a step size of $\alpha = 0.01$, such that for each action-reward observation tuple $(a_t = i, r_{t+1})$, $\bar{q}_i \Leftarrow \alpha(r_{t+1} - \bar{q}_i)$, with initial estimates being uniformly distributed in $[0, 1]$. In the other two cases, the complete distribution of q_i was maintained via a population $\{p_i^k\}_{k=0}^K$ of point estimates, with $K = 16$. Each point estimate in the population was maintained in the same manner as the single point estimates in the ϵ-greedy approach. Sampling actions was performed by sampling uniformly from the members of the population for each action.

The results for two different bandit tasks, one with 16 and the other with 128 arms, averaged over 1,000 runs, are summarised in Fig. 6. For each run, the expected reward of each bandit was sampled uniformly from $[0, 1]$. As can be seen from the figure, the ϵ-greedy approach performs relatively well when used with reasonable first initial estimates. The sampling greedy approach, while having the same complexity, appears to perform better asymptotically. More importantly, Alg. 2 exhibits better long-term versus short-term performance when the effective reward horizon is increased as $\gamma \to 1$.

Algorithm 3 Sampling-greedy

$a \Leftarrow i$ with probability $P(a = i) = P(q_i > q_j)\ \forall j \neq i$

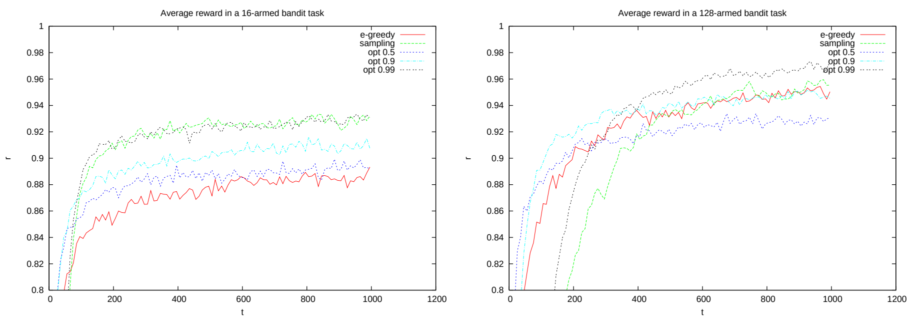

Fig. 1. Average reward in an multi-armed bandit task averaged over 1,000 experiments, smoothed with a moving average over 10 time-steps. Results are shown for ϵ-greedy (**e-greedy**), sampling-greedy (**sampling**) and Alg. 2 (**opt**) with $\gamma \in \{0.5, 0.9, 0.99\}$.

7 Discussion and Conclusion

This paper has presented a formulation of an optimal exploration-exploitation threshold for in a n-armed bandit task, which links the need for exploration to the effective reward horizon and model uncertainty. Additionally, a practical algorithm, based on an optimistic bound on the value of exploration, is introduced. Experimental results show that this algorithm exhibits the expected long-term versus short-term performance trade-off when the effective reward horizon is increased.

While the above formulation fits well within a reinforcement learning framework, other useful formulations may exist. In *budgeted learning*, any exploratory action results in a fixed cost. Such a formulation is used in [10] for the bandit problem (see also [11] for the active learning case). Then the problem essentially becomes that of how to best sample from actions in the next T moves such that the expected return of the optimal policy *after* T moves is maximised and corresponds to $g(k) = 0 \; \forall k < T$ in the framework presented in this paper. A further alternative, described in [6], is to stop exploring those parts of the state-action space which lead to sub-optimal returns with high probability.

When a distribution or a confidence interval is available for expected returns, it is common to use the optimistic side of the confidence interval for action selection [12]. This practice can be partially justified through the framework presented herein, or alternatively, through considering maximising the expected information to be gained by exploration, as proposed by [13]. In a similar manner, other methods which represent uncertainty as a simple additive factor to the normal expected reward estimates, acquire further meaning when viewed through a statistical decision making framework. For example the Dyna-Q+ algorithm (see [7] chap. 9) includes a slowly increasing *exploration bonus* for state-action pairs which have not been recently explored. From a statistical viewpoint, the exploration bonus corresponds to a model of a non-stationary world, where uncertainty about past experiences increases with elapsed time.

In general, the conditions defined in Sec. 3 require maintaining some type of belief distribution over the expected return of actions. A natural choice for this would be to use a fully analytical Bayesian framework. Unfortunately this makes it more difficult to calculate $P(q_i > d)$, so it might be better to consider simple numerical approaches from the outset. We have previously considered some simple such estimates in [14], where we relied on estimating the gradient of the expected return with respect to the parameters. The estimated gradient was then used as a measure of uncertainty. Further research on the use of population-based methods for explicitly representing a distribution of estimates is currently under way.

References

1. Wald, A.: Sequential Analysis. John Wiley & Sons (1947) Republished by Dover in 2004.
2. DeGroot, M.H.: Optimal Statistical Decisions. John Wiley & Sons (1970) Republished in 2004.
3. Bellman, R.E.: A problem in the sequential design of experiments. Sankhya **16** (1957) 221–229
4. Mannor, S., Tsitsiklis, J.N.: The sample complexity of exploration in the multi-armed bandit problem. Journal of Machine Learning Research **5** (2004) 623–648
5. Dearden, R., Friedman, N., Russell, S.J.: Bayesian Q-learning. In: AAAI/IAAI. (1998) 761–768
6. Even-Dar, E., Mannor, S., Mansour, Y.: Action elimination and stopping conditions for the multi-armed and reinforcement learning problems. Journal of Machine Learning Research (2006) to appear.
7. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press (1998)
8. Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., Knuth, D.E.: On the lambert W function. Advances in Computational Mathematics **5** (1996) 329–359
9. Sutton, R.S., Precup, D., Singh, S.P.: Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence **112**(1-2) (1999) 181–211
10. Madani, O., Lizotte, D.J., Greiner, R.: The budgeted multi-armed bandit problem. In: Learning Theory: 17th Annual Conference on earning Theory, COLT 2004. Volume 3120 of Lecture Notes in Computer Science., Springer-Verlag (2004) 643–645
11. Madani, O., Lizotte, D.J., Greiner, R.: Active model selection. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada, AUAI Press, Arlington, Virginia (2004) 357–365
12. Auer, P.: Models for trading exploration and exploitation using upper confidence bounds. In: PASCAL workshop on principled methods of trading exploration and exploitation, PASCAL Network (2005)
13. Bernardo, J.M.: Expected information as expected utility. In: The Annals of Statistics. Volume 7., Institute of Mathematical Statistics (1979) 686–690
14. Dimitrakakis, C., Bengio, S.: Gradient estimates of return. IDIAP-RR 05-29, IDIAP (2005)