JMLR: Workshop and Conference Proceedings 24:1–9, 2012 10th European Workshop on Reinforcement Learning

Learning Exploration/Exploitation Strategies for Single Trajectory Reinforcement Learning

Michael Castronovo MICHAEL.CASTRONOVO@STUDENT.ULG.AC.BE
Francis Maes FRANCIS.MAES@ULG.AC.BE
Raphael Fonteneau RAPHAEL.FONTENEAU@ULG.AC.BE
Damien Ernst DERNST@ULG.AC.BE
Department of Electrical Engineering and Computer Science, University of Liège, BELGIUM

Editor: Marc Peter Deisenroth, Csaba Szepesvári, Jan Peters

Abstract

We consider the problem of learning high-performance Exploration/Exploitation (E/E) strategies for finite Markov Decision Processes (MDPs) when the MDP to be controlled is supposed to be drawn from a known probability distribution $p_{\mathcal{M}}(\cdot)$. The performance criterion is the sum of discounted rewards collected by the E/E strategy over an infinite length trajectory. We propose an approach for solving this problem that works by considering a rich set of candidate E/E strategies and by looking for the one that gives the best average performances on MDPs drawn according to $p_{\mathcal{M}}(\cdot)$. As candidate E/E strategies, we consider index-based strategies parametrized by small formulas combining variables that include the estimated reward function, the number of times each transition has occurred and the optimal value functions $\hat{V}$ and $\hat{Q}$ of the estimated MDP (obtained through value iteration). The search for the best formula is formalized as a multi-armed bandit problem, each arm being associated with a formula. We experimentally compare the performances of the approach with R-MAX as well as with ϵ-GREEDY strategies and the results are promising.

Keywords: Reinforcement Learning, Exploration/Exploitation dilemma, Formula discovery

1. Introduction

Most Reinforcement Learning (RL) techniques focus on determining high-performance policies maximizing the expected discounted sum of rewards to come using several episodes. The quality of such a learning process is often evaluated through the performances of the final policy regardless of rewards that have been gathered during learning. Some approaches have been proposed to take these rewards into account by minimizing the undiscounted regret (Kearns and Singh (2002); Brafman and Tennenholtz (2002); Auer and Ortner (2007); Jaksch et al. (2010)), but RL algorithms have troubles solving the *original RL problem* of maximizing the expected discounted return over a single trajectory. This problem is almost intractable in the general case because the discounted nature of the regret makes early mistakes - often due to hazardous exploration - almost impossible to recover from. Roughly speaking, the agent needs to learn very fast in one pass. One of the best solutions to face this Exploration/Exploitation (E/E) dilemma is the R-MAX algorithm (Brafman and Tennenholtz (2002)) which combines model learning and dynamic programming with the "optimism in the face of uncertainty" principle. However, except in the case where the underlying Markov Decision Problem (MDP) comes with a small number of states and a discount factor very close to 1 (which corresponds to giving more chance to recover from bad initial decisions), the performance of R-max is still very far from the optimal (more details in Section 5).

In this paper, we assume some prior knowledge about the targeted class of MDPs, expressed in the form of a probability distribution over a set of MDPs. We propose a scheme for learning E/E strategies that makes use of this probability distribution to sample training MDPs. Note that this assumption is quite realistic, since before truly interacting with the MDP, it is often possible to have some prior knowledge concerning the number of states and actions of the MDP and/or the way rewards and transitions are distributed.

To instantiate our learning approach, we consider a rich set of candidate E/E strategies built around parametrized index-functions. Given the current state, such index-functions rely on all transitions observed so far to compute E/E scores associated to each possible action. The corresponding E/E strategies work by selecting actions that maximize these scores. Since most previous RL algorithms make use of small formulas to solve the E/E dilemma, we focus on the class of index-functions that can be described by a large set of such small formulas. We construct our E/E formulas with variables including the estimated reward function of the MDP (obtained from observations), the number of times each transition has occurred and the estimated optimal value functions $\hat{V}$ and $\hat{Q}$ (computed through off-line value iteration) associated with the estimated MDP. We then formalize the search for an optimal formula within that space as a multi-armed bandit problem, each formula being associated to an arm.

Since it assumes some prior knowledge given in the form of a probability distribution over possible underlying MDPs, our approach is related to Bayesian RL (BRL) approaches (Poupart et al. (2006); Asmuth et al. (2009)) that address the E/E trade-off by (i) assuming a prior over possible MDP models and (ii) maintaining - from observations - a posterior probability distribution (i.e., "refining the prior"). In other words, the prior is used to reduce the number of samples required to construct a good estimate of the underlying MDP and the E/E strategy itself is chosen a priori following Bayesian principles and does not depend on the targeted class of MDPs. Our approach is specific in the sense that the prior is not used for better estimating the underlying MDP but rather for identifying the best E/E strategy for a given class of targeted MDPs, among a large class of diverse strategies. We therefore follow the work of Maes et al. (2012), which already proposed to learn E/E strategies in the context of multi-armed bandit problems, which can be seen as state-less MDPs.

This paper is organized as follows. Section 2 formalizes the E/E strategy learning problem. Section 3 describes the space of formula-based E/E strategies that we consider in this paper. Section 4 details our algorithm for efficiently learning formula-based E/E strategies. Our approach is illustrated and empirically compared with R-max as well as with ϵ-Greedy strategies in Section 5. Finally, Section 6 concludes.

2. Background

Let $M = (\mathcal{S}, \mathcal{A}, p_{M,f}(\cdot), \rho_M, p_{M,0}(\cdot), \gamma)$ be a MDP. $\mathcal{S} = \left\{s^{(1)}, \ldots, s^{(n_{\mathcal{S}})}\right\}$ is its state space and $\mathcal{A} = \left\{a^{(1)}, \ldots, a^{(n_{\mathcal{A}})}\right\}$ its action space. When the MDP is in state s_t at time t and action a_t is selected, the MDP moves to a next state s_{t+1} drawn according to the probability distribution $p_{M,f}(\cdot|s_t, a_t)$. A deterministic instantaneous scalar reward $r_t = \rho_M(s_t, a_t, s_{t+1})$ is associated with the stochastic transition (s_t, a_t, s_{t+1}).

$H_t = [s_0, a_0, r_0, \ldots, s_t, a_t, r_t]$ is a vector that gathers the history over the first t steps and we denote by $\mathcal{H}$ the set of all possible histories of any length. An exploration / exploitation (E/E) strategy is a stochastic algorithm π that, given the current state s_t, processes at time t the vector H_{t-1} to select an action $a_t \in \mathcal{A}$: $a_t \sim \pi(H_{t-1}, s_t)$. Given the probability distribution over initial states $p_{M,0}(\cdot)$, the performance/return of a given E/E strategy π with respect to the MDP M can be defined as: $J^\pi_M = \underset{p_{M,0}(\cdot), p_{M,f}(\cdot)}{\mathbb{E}} [\mathcal{R}^\pi_M(s_0)]$ where $\mathcal{R}^\pi_M(s_0)$ is the stochastic discounted return of the E/E strategy π when starting from the state s_0. This return is defined as:

$$\mathcal{R}^\pi_M(s_0) = \sum_{t=0}^{\infty} \gamma^t r_t \ ,$$

where $r_t = \rho_M(s_t, \pi(H_{t-1}, s_t), s_{t+1})$ and $s_{t+1} \sim p_{M,f}(.|s_t, \pi(H_{t-1}, s_t))$ $\forall t \in \mathbb{N}$ and where the discount factor γ belongs to $[0, 1)$. Let $p_{\mathcal{M}}(\cdot)$ be a probability distribution over MDPs, from which we assume that the actual underlying MDP M is drawn. Our goal is to learn a high performance finite E/E strategy π given the prior $p_{\mathcal{M}}(\cdot)$, i.e. an E/E strategy that maximizes the following criterion:

$$J^\pi = \underset{M' \sim p_{\mathcal{M}}(\cdot)}{\mathbb{E}} [J^\pi_{M'}] \ . \tag{1}$$

3. Formula-based E/E strategies

In this section, we describe the set of E/E strategies that are considered in this paper.

3.1. Index-based E/E strategies

Index-based E/E strategies are implicitly defined by maximizing history-dependent state-action index functions. Formally, we call a history-dependent state-action index function any mapping $I : \mathcal{H} \times \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Given such an index function I, a decision can be taken at time t in the state $s_t \in \mathcal{S}$ by drawing an optimal action according to I: $\pi(H_{t-1}, s_t) \in \underset{a \in \mathcal{A}}{\arg\max}\ I(H_{t-1}, s_t, a)$[1]. Such a procedure has already been vastly used in the particular case where the index function is an estimate of the action-value function, eventually randomized using $\epsilon-$greedy or Boltzmann exploration, as in Q-Learning (Watkins and Dayan (1992)).

[1] Ties are broken randomly in our experiments.

3.2. Formula-based E/E strategies

We consider in this paper index functions that are given in the form of small, closed-form formulas. This leads to a very rich set of candidate E/E strategies that have the advantage of being easily interpretable by humans. Formally, a formula $F \in \mathbb{F}$ is:

- either a binary expression $F = B(F', F'')$, where B belongs to a set of binary operators $\mathbb{B}$ and F' and F'' are also formulas from $\mathbb{F}$,
- or a unary expression $F = U(F')$ where U belongs to a set of unary operators $\mathbb{U}$ and $F' \in \mathbb{F}$,
- or an atomic variable $F = V$, where V belongs to a set of variables $\mathbb{V}$ depending on the history H_{t-1}, the state s_t and the action a,
- or a constant $F = C$, where C belongs to a set of constants $\mathbb{C}$.

Since it is high dimensional data of variable length, the history H_{t-1} is non-trivial to use directly inside E/E index-functions. We proceed as follows to transform the information contained in H_{t-1} into a small set of relevant variables. We first compute an estimated model of the MDP $\hat{M}$ that differs from the original M due to the fact that the transition probabilities and the reward function are not known and need to be learned from the history H_{t-1}. Let $\hat{P}(s, a, s')$ and $\hat{\rho}(s, a)$ be the transition probabilities and the reward function of this estimated model. $\hat{P}(s, a, s')$ is learned by computing the empirical frequency of jumping to state s' when taking action a in state s and $\hat{\rho}(s, a)$ is learned by computing the empirical mean reward associated to all transitions originating from (s, a)[2]. Given the estimated MDP, we run a value iteration algorithm to compute the estimated optimal value functions $\hat{V}(\cdot)$ and $\hat{Q}(\cdot, \cdot)$. Our set of variables is then defined as: $\mathbb{V} = \left\{ \hat{\rho}(s_t, a), N(s_t, a), \hat{Q}(s_t, a), \hat{V}(s_t), t, \gamma^t \right\}$ where $N(s, a)$ is the number of times a transition starting from (s, a) has been observed in H_{t-1}.

We consider a set of operators and constants that provides a good compromise between high expressiveness and low cardinality of $\mathbb{F}$. The set of binary operators $\mathbb{B}$ includes the four elementary mathematical operations and the min and max operators: $\mathbb{B} = \{+, -, \times, \div, \min, \max\}$. The set of unary operators $\mathbb{U}$ contains the square root, the logarithm and the absolute value: $\mathbb{U} = \left\{ \sqrt{\cdot}, \ln(\cdot), |\cdot| \right\}$. The set of constants is: $\mathbb{C} = \{1, 2, 3, 5, 7\}$.

In the following, we denote by π^F the E/E strategy induced by formula F:

$$\pi^F(H_{t-1}, s_t) \in \underset{a \in \mathcal{A}}{\arg\max}\ F\left(\hat{\rho}(s_t, a), N(s_t, a), \hat{Q}(s_t, a), \hat{V}(s_t), t, \gamma^t \right)$$

We denote by $|F|$ the description length of the formula F, i.e. the total number of operators, constants and variables occurring in F. Let K be a maximal formula length. We denote by $\mathbb{F}^K$ the set of formulas whose length is not greater than K. This defines our so-called set of small formulas.

[2] If a pair (s, a) has not been visited, we consider the following default values: $\hat{\rho}(s, a) = 0$, $\hat{P}(s, a, s) = 1$ and $\hat{P}(s, a, s') = 0, \forall s' \neq s$.

4. Finding a high-performance formula-based E/E strategy for a given class of MDPs

We look for a formula F^* whose corresponding E/E strategy is specifically efficient for the subclass of MDPs implicitly defined by the probability distribution $p_{\mathcal{M}}(\cdot)$. We first describe a procedure for accelerating the search in the space $\mathbb{F}^K$ by eliminating equivalent formulas in Section 4.1. We then describe our optimization scheme for finding a high-performance E/E strategy in Section 4.2.

4.1. Reducing $\mathbb{F}^K$

Notice first that several formulas $\mathbb{F}^K$ can lead to the same policy. All formulas that rank all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ in the same order define the same policy. We partition the set $\mathbb{F}^K$ into equivalence classes, two formulas being equivalent if and only if they lead to the same policy. For each equivalence class, we then consider one member of minimal length, and we gather all those minimal members into a set $\bar{\mathbb{F}}^K$.

Computing the set $\bar{\mathbb{F}}^K$ is not trivial: given a formula, equivalent formulas can be obtained through commutativity, associativity, operator-specific rules and through any increasing transformation. We thus propose to approximately discriminate between formulas by comparing how they rank (in terms of values returned by the formula) a set of d random samples of the variables $\hat{\rho}(\cdot,\cdot), N(\cdot,\cdot), \hat{Q}(\cdot,\cdot), \hat{V}(\cdot), t, \gamma^t$. More formally, the procedure is the following:

- we first build $\mathbb{F}^K$, the space of all formulas such that $|F| \leq K$;
- for $i = 1 \ldots d$, we uniformly draw (within their respective domains) some random realizations of the variables $\hat{\rho}(\cdot,\cdot), N(\cdot,\cdot), \hat{Q}(\cdot,\cdot), \hat{V}(\cdot), t, \gamma^t$ that we concatenate into a vector Θ_i;
- we cluster all formulas from $\mathbb{F}^K$ according to the following rule: two formulas F and F' belong to the same cluster if and only if they rank all the Θ_i points in the same order, i.e.: $\forall i, j \in \{1, \ldots, d\}, i \neq j, F(\Theta_i) \geq \ F(\Theta_j) \iff F'(\Theta_i) \geq \ F'(\Theta_j)$. Formulas leading to invalid index functions (caused for instance by division by zero or logarithm of negative values) are discarded;
- among each cluster, we select one formula of minimal length;
- we gather all the selected minimal length formulas into an approximated reduced set of formulas $\tilde{\mathbb{F}}^K$.

In the following, we denote by N the cardinality of the approximate set of formulas $\tilde{\mathbb{F}}^K = \{F_1, \ldots, F_N\}$.

4.2. Finding a high-performance formula

A naive approach for determining a high-performance formula $F^* \in \tilde{\mathbb{F}}^K$ would be to perform Monte-Carlo simulations for all candidate formulas in $\tilde{\mathbb{F}}^K$. Such an approach could reveal itself to be time-inefficient in case of spaces $\tilde{\mathbb{F}}^K$ of large cardinality.

We propose instead to formalize the problem of finding a high-performance formula-based E/E strategy in $\tilde{\mathbb{F}}^K$ as a $N-$armed bandit problem. To each formula $F_n \in \tilde{\mathbb{F}}^K$ $(n \in \{1, \ldots, N\})$, we associate an arm. Pulling the arm n consists first in randomly drawing a MDP M according to $p_{\mathcal{M}}(\cdot)$ and an initial state s_0 for this MDP according to $p_{M,0}(\cdot)$.

Afterwards, an episode starting from this initial state is generated with the E/E strategy π^{F_n} until a truncated time horizon T. This leads to a reward associated to arm n whose value is the discounted return $\mathcal{R}^{\pi}_{M}(s_0)$ observed during the episode. The purpose of multi-armed bandit algorithms is here to process the sequence of such observed rewards to select in a smart way the next arm to be played so that when the budget of pulls has been exhausted, one (or several) high-quality formula(s) can be identified.

Multi-armed bandit problems have been vastly studied, and several algorithms have been proposed, such as for instance all UCB-type algorithms (Auer et al. (2002); Audibert et al. (2007)). New approaches have also recently been proposed for identifying automatically empirically efficient algorithms for playing multi-armed bandit problems (Maes et al. (2011)).

5. Experimental results

In this section, we empirically analyze our approach on a specific class of random MDPs defined hereafter.

Random MDPs. MDPs generated by our prior $p_{\mathcal{M}}(\cdot)$ have $n_{\mathcal{S}} = 20$ states and $n_{\mathcal{A}} = 5$ actions. When drawing a MDP according to this prior, the following procedure is called for generating $p_{M,f}(\cdot)$ and $\rho_M(\cdot,\cdot,\cdot)$. For every state-action pair (s,a) : (i) it randomly selects 10% of the states to form a set of successor states $Succ(s,a) \subset \mathcal{S}$ (ii) it sets $p_{M,f}(s'|s,a) = 0$ for each $s' \in \mathcal{S} \setminus Succ(s,a)$ (iii) for each $s' \in Succ(s,a)$, it draws a number $N(s')$ at random in $[0,1]$ and sets $p_{M,f}(s'|s,a) = \frac{N(s')}{\sum_{s'' \in Succ(s,a)} N(s'')}$ (iv) for each $s' \in Succ(s,a)$, it sets $\rho_M(s,a,s')$ equal to a number chosen at random in $[0,1]$ with a 0.1 probability and to zero otherwise. The distribution $p_{M,0}(\cdot)$ of initial states is chosen uniform over $\mathcal{S}$. The value of γ is equal to 0.995.

Learning protocol. In our experiments, we consider a maximal formula length of $K = 5$ and use $d = 1000$ samples to discriminate between formulas, which leads to a total number of candidate E/E strategies $N = 3834$. For solving the multi-armed bandit problem described in Section 4.2, we use an Upper Confidence Bound (UCB) algorithm (Auer et al. (2002)). The total budget allocated to the search of a high-performance policy is set to $T_b = 10^6$. We use a truncated optimization horizon $T = \log_{\gamma}((1-\gamma)\delta)$ for estimating the stochastic discounted return of an E/E strategy where $\delta = 0.001$ is the chosen precision (which is also used as stopping condition in the off-line value iteration algorithm for computing $\hat{Q}$ and $\hat{V}$). At the end of the T_b plays, the five E/E strategies that have the highest empirical return mean are returned.

Baselines. Our first baseline, the Optimal strategy, consists in using for each test MDP, a corresponding optimal policy. The next baselines, the Random and Greedy strategies perform pure exploration and pure exploitation, respectively. The Greedy strategy is equivalent to an index-based E/E strategy with formula $\hat{Q}(s,a)$. The last two baselines are classical E/E strategies whose parameters have been tuned so as to give the best performances on MDPs drawn from $p_{\mathcal{M}}(\cdot)$: ϵ-Greedy and R-max. For ϵ-Greedy, the best value we found was $\epsilon = 0$ in which case it behaves as the Greedy strategy. This confirms that hazardous exploration is particularly harmful in the context of single trajectory RL with discounted return. Consistently with this result, we observed that R-max works at its best when it performs the least exploration ($m = 1$).

Baselines		Learned strategies	
Name	J^π	Formula	J^π
Optimal	65.3	$(N(s,a) \times \hat{Q}(s,a)) - N(s,a)$	30.3
Random	10.1	$\max(1, (N(s,a) \times \hat{Q}(s,a)))$	22.6
Greedy	20.0	$\hat{Q}(s,a)$ (= Greedy)	20.0
ϵ-Greedy($\epsilon = 0$)	20.0	$\min(\gamma^t, (\hat{Q}(s,a) - \hat{V}(s)))$	19.4
R-max ($m = 1$)	27.7	$\min(\hat{\rho}(s,a), (\hat{Q}(s,a) - \hat{V}(s)))$	19.4

Table 1: Performance of the top-5 learned strategies with respect to baseline strategies.

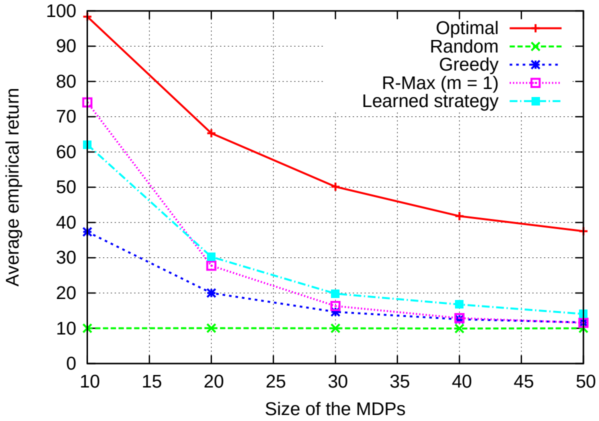

Figure 1: Performances of the learned and the baseline strategies for different distributions of MDPs that differ by the size of the MDPs belonging to their support.

Results. Table 1 reports the mean empirical returns obtained by the E/E strategies on a set of 2000 test MDPs drawn from $p_{\mathcal{M}}(\cdot)$. Note that these MDPs are different from those used during learning and tuning. As we can see, the best E/E strategy that has been learned performs better than all baselines (except the Optimal), including the state-of-the-art approach R-max.

We may wonder to what extent the E/E strategies found by our learning procedure would perform well on MDPs which are not generated by $p_{\mathcal{M}}(\cdot)$. As a preliminary step to answer this question, we have evaluated the mean return of our policies on sets of 2000 MDPs drawn from slightly different distributions as the one used for learning: we changed the number of states $n_{\mathcal{S}}$ to different values in $\{10, 20, \ldots, 50\}$. The results are reported in Figure 1. We observe that, except in the case $n_{\mathcal{S}} = 10$, our best E/E strategy still performs better than the R-max and the ϵ-Greedy strategies tuned on the original distribution $p_{\mathcal{M}}(\cdot)$ that generates MDPs with 20 states. We also observe that for larger values of $n_{\mathcal{S}}$, the performances of R-max become very close to those of Greedy, whereas the performances of our best E/E strategy remain clearly above. Investigating why this formula performs well is left for future work, but we notice that it is analog to the formula $t_k(r_k - C)$ that was automatically discovered as being well-performing in the context of multi-armed bandit problems (Maes et al. (2011)).

6. Conclusions

In this paper, we have proposed an approach for learning E/E strategies for MDPs when the MDP to be controlled is supposed to be drawn from a known probability distribution $p_{\mathcal{M}}(\cdot)$. The strategies are learned from a set of training MDPs (drawn from $p_{\mathcal{M}}(\cdot)$) whose size depends on the computational budget allocated to the learning phase. Our results show that the learned strategies perform very well on test problems generated from the same distribution. In particular, they outperform on these problems R-max and ϵ-Greedy policies. Interestingly, the strategies also generalize well to MDPs that do not belong to the support of $p_{\mathcal{M}}(\cdot)$. This is demonstrated by the good results obtained on MDPs having a larger number of states than those belonging to $p_{\mathcal{M}}(\cdot)$'s support.

These encouraging results suggest several future research direction. First, it would be interesting to better study the generalization performances of our approach either theoretically or empirically. Second, we believe that our approach could still be improved by considering richer sets of formulas w.r.t. the length of the formulas and the number of variables extracted from the history. Finally, it would be worth investigating ways to improve the optimization procedure upon which our learning approach is based so as to be able to deal with spaces of candidate E/E strategies that are so large that even running once every strategy on a single training problem would be impossible.

References

J. Asmuth, L. Li, M.L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26. AUAI Press, 2009.

J.Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165. Springer, 2007.

P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 49. The MIT Press, 2007.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

R.I. Brafman and M. Tennenholtz. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2002.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

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

F. Maes, L. Wehenkel, and D. Ernst. Automatic discovery of ranking formulas for playing with multi-armed bandits. In *9th European workshop on reinforcement learning*, Athens, Greece, September 2011.

F. Maes, L. Wehenkel, and D. Ernst. Learning to play K-armed bandit problems. In *International Conference on Agents and Artificial Intelligence*, Vilamoura, Algarve, Portugal, February 2012.

P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704. ACM, 2006.

C.J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):179–192, 1992.