Influence of the context of a Reinforcement Learning Technique on the learning performances - A case study

Frédéric Davesne
LPPA
CNRS UMR 7124 - Collège de France
11, Place Marcelin-Berthelot
75005 Paris - France
frederic.davesne@college-de-france.fr

Claude Barret
LSC
CNRS FRE 2494 - University of Evry
40, Rue du Pelvoux
91020 Evry Cedex - France
claude.barret@iup.univ-evry.fr

ABSTRACT

Statistical learning methods select the model that statistically best fit the data, given a cost function. In this case, learning means finding out a set of internal parameters of the model that minimize (or maximize) the cost function. As an example of such a procedure, reinforcement learning techniques (RLT) may be used in robotics to find the best mapping between sensors and effectors to achieve a goal.
A lot of practical issues have been already pointed out to apply RLT in real robotics, and some solutions have been investigated. However, an underlying issue, which is critical for the reliability of the task accomplished by the robot, is the adequacy of the a priori knowledge (design of the states, value of the temperature parameter) used by the RLT with the physical properties of the robot, in order to achieve the goal defined by the experimenter. We call it Context Quality (CQ). Some work has pointed out that bad CQ may lead to poor learning results, but CQ in itself was not really quantified.
In this paper, we suggest that the entropy measure taken from the Information Theory is well suited to quantify CQ and to predict the quality of the results obtained by the learning process. Taking the Cart Pole Balancing benchmark, we show that there exists a strong relation between our CQ measure and the performance of the RLT, that is to say the viability duration of the cart/pole. In particular, we investigate the influence of the noisiness of the inputs and the design of the states. In the first case, we show that CQ is linked to performance of recognition of the input states by the system. Moreover, we propose an statistical explanatory model of the influence of CQ on the RLT performance.

KEY WORDS
Machine Learning, Context Quality, State Design Testing, Shannon Entropy.

1 Introduction

1.1 Framework

Reinforcement Learning (RL) is an optimization tool, derived from Dynamic Programming [15]. It permits to learn the local association between input and output data in order to produce a "good" sequence of outputs to achieve a goal. Typically, the input is a set of states (finite or infinite) and the output is a set of actions (finite or infinite) that the system may perform. RL is locally directed by a (coarse) signal - the reinforcement value - which establishes a distance to the goal. Hence, RL permits to integrate the reinforcement values trough time in order to build a cost function that measures the quality of each possible action, given a state.

Theoretical results exist for some Reinforcement Learning Techniques (RLT): Dayan has shown convergence properties of Q-Learning [6] (finite set of states) and Munos extended the former result to the continuous case [11].

RL has led to numerous successful applications, in particular for "pure" optimization problems, in which the states are exactly known. Some good results have been obtained in the area of command (the cart/pole balancing problem was the first well-known application [1]), simulated robotics [10]. But it has been experienced that even a small amount of noise may produce an unstable learning, which leads to poor results. Pendrith studied the impact of noise on the RLT performance [13], [12].

The fact that the decision problem becomes non-markovian is the main reason for explaining the lack of performance of RLT when input date are noisy. It is true that, in this case, convergence to an optimal policy is not theoretically guarantied. A practical solution may consist in applying a low-pass filter to the input data to smoothen them, or to utilize variation of Q-Learning that permits to cope with imprecise input data: Glorennec has mixed Fuzzy Logic and Reinforcement Learning [7]. Another solution, which has been explored in "pure" optimization problems, is to suppose that states are not directly observable but may be deduced from the input data: POMDP techniques are based on this idea [9]. However, this idea is not really ap-

plicable in real robotics because the states are not really hidden to the observer: there is a difficulty to discriminate a state from another.

The non-Markovian case may be the result of two issues:

- a state in itself is precisely known given the input data, but the design of the set of states is not compatible with the actions and the goal to be achieved.
- a state is not precisely known, given the input data

We call theses issues *contextual issues* because RLT are not supposed to solve them, although they clearly impact the learning performances. Real robotics sums up the two difficulties, because data are noisy and the experimenter designs the states by using his own perception of the environment of the robot, which may be incompatible with the perception capabilities of the robot: this was depicted by Harnad as the Symbol Grounding Problem [8].

1.2 Focus

The impact of the context on the performance of RLT has not been really studied. In fact, in the case of Cart Pole Balancing, performances obtained by different RLT may vary considerably. We raise the following question: is this difference due to the RLT in itself or to the context that goes with the RLT ? We make the general postulate that *the Context Quality (CQ)* has a deep impact on the learning results.

If this postulate is true, knowing CQ before the learning process may permit to predict the performance obtained by the learning phase. Moreover, if CQ could be quantified, it would be possible to construct the context of RLT in order to maximize (or minimize) it. A full study of this issue includes:

- a specification of a CQ measure that is influenced by all the parameters or algorithms that are not modified by RLT.
- a method to build an Ideal Context, that maximize (or minimize) CQ

In this paper, we will focus on the study a CQ measure which values are influenced by the input data/state association process, including:

- the a priori design of the states
- the mechanism which associate raw input data to a particular state

In the following, we will call this process the State Recognition Process (SRP).

The CQ measure we have chosen is based on the Shannon entropy. It is linked with two kind of informations:

1. to what extent is it possible to discriminated states using the association mechanism ?
2. to what extent is it possible to predict the future state knowing specific action and raw data ?

The best-case scenario (which minimize CQ) is the labyrinth benchmark in which each input data is perfectly associated to a unique state (the discrimination between states is maximum) and where a future state may be perfectly predicted, knowing the input data and an action. So, in our mind, CQ is related to two issues: state recognition (SR) and future state prediction (SP). The best SR and SP are performed, the less CQ is.

The Markovian case may be seen as a case where SR is well done and SP may be not well accomplished. Given a state, the worst possibility here consists on having the same probability to move from this state to all other states by using an action. For the best case, all but one of the transition probabilities are 0 and one is 1: here, the transition is deterministic.

Our CQ definition may appear to be unrealistic, because the set of states linked with an ideal context is ruled by deterministic transitions and it is always possible to know very accurately in which state the system is: it is similar to the Turing machine case. Even a simple application like the Cart Pole Balancing designed by Barto et al. [1] is not associated to an ideal context (see par. 2.3) (SP cannot be precisely done, with the state specification of Barto et al.): nevertheless, the results are good (the cart/pole is successfully balanced for at least 100000 consecutive steps).

We claim that the design of states is critical and must be done regarding CQ. In this article, we show, in particular, that the goodness of the results obtained for the Cart Pole Balancing problem must be taken carefully: if we fix a much more larger threshold to decide that a learning trial has succeeded, let's say 100 million consecutive steps, we remark that the system is barely able to achieve its goal (see par. 3.2). That means the design of the states, like it was done by Barto et al., do not permit to produce a perfectly reliable action policy. We suggest that the failures are not due to the RLT in itself, but to the context of RLT, even if raw input data are not noisy.

Another question that may be asked is about the necessity of using RLT within an nearly-ideal context. If the transition probabilities from a state to another are near 0 or 1, is it interesting to use a statistical tool ? Few years ago, we developed a specific algorithm, called Constraint based Learning (CbM), which is applicable in the case where CQ is quite small. The description of CbM is out of the topic of this article. However, one may refer to [5] and [4] to have an application of CbM for navigation tasks of a Khepera robot. Theoretical results, in a near-ideal context, concerning the convergence of CbM and its incremental characteristics has been proved in [3]. Results from the labyrinth benchmark have shown that CbM is considerably faster than Q-Learning and one of its improvements $Q(\lambda)$.

1.3 Experimental environment

1.3.1 Design of the experiment

We will utilize the Cart Pole Balancing benchmark. Four input variables are considered: the cart position and speed (namely x and $\dot{x}$), and the pole position and speed (namely θ and $\dot{\theta}$). We will use the same SRP as in [1], but will add some artificial noise to the raw input data, so that the output state of the SRP is influenced by this noise. We will take into account three types of noise which will be applied on θ:

- (GN) A zero-mean Gaussian noise, with standard deviation σ.
- (OLN) Outliers produced with a rate r_o. Outliers are values taken from a Uniform Law into the interval [-0.2 rad, 0.2 rad]
- (RSN) The output state of SRP is chosen randomly with a Uniform Law on the set of states, with a rate r_r

1.3.2 Learning procedure

The RLT we have chosen is $Q(\lambda)$ [14], derived from Q-Learning. The learning phase consists of 2000 trials. We decide to fix the number of consecutive steps associated to a success of a learning trial to a much higher value than in Barto et al.: 100 million steps. This permits to test the reliability of the action policy found by RLT, given a precise SRP. We want to prove that the SRP designed in Barto et al. do not permit to achieve our required performance. For each trial of the learning phase, the initial raw input data corresponding to $(x, \dot{x}, \theta, \dot{\theta})$ is chosen randomly (Uniform Law) in the hypercube [-0.8,0.8]x[-0.5,0.5]x[-6,6]x[-0.87,0.87]. We use a pseudo-exhaustive method to fix the choice of action policy: the action linked to the best Q-value is chosen with a probability P. It is important to stress that P is a constant: we have not managed to balance successfully the cart/pole for 100 million steps with a decreasing P over time.

2 The Context Quality measure

2.1 Choice of the Context Quality measure

We have chosen to measure the information transmitted by the change from one state to another, using a precise action. A lot of measuring tools may be suitable. Bouchon explains that the choice between them depends on the nature of the information, which can be parted into two classes [2]:

- observation information, which permits to evaluate the precision of the input data.
- exploitation information, which permits to take a decision

The two kind of informations are mixed together in our case: the result of the execution of an action at time t may be uncertain, because we do not know accurately the state at t (due to noisy input data) and because we cannot predict the resulting state at time t+1. The Entropic Model Theory considers two kinds of models [2]:

1. entropic models of type 1, which deals with the uncertainty due to the tool used for getting the observations
2. entropic models of type 2, which deals with the impreciseness of the observations

We do not want to evaluate the impreciseness of the input data, but the resulting uncertainty on the knowledge we have about the state at time t and t+1. So, we are in the first case and may use Shannon entropy, Hartley information or Kullback-Leiber information. We have chosen the Shannon entropy.

2.2 Notations and specification

We consider that a RLT utilizes a finite set of n states $e_1, e_2, \ldots, e_n$ and a finite set of q actions $a_1, a_2, \ldots, a_q$. The states e_i are deduced from raw input data. We also consider a set of $n.q$ transitory states $e_{i,k}$ which denotes that action a_k has been performed from state e_i. The probability for the system to jump from state e_i to state $e_j, j \neq i$ by using action a_k is $p_{i,k,j}$. This term corresponds to the transition probability in the Q-Learning algorithm. It is important to notice that the states e_i or $e_{i,k}$ are not the "true" states, but the output of an algorithm which inputs are raw data. This algorithm performs a SRP, which belongs to the context of the RLT.

Now, we specify a first term for the measure of CQ for one state e_i. First, we create a term $H(e_i, a_k)$ that characterizes the uncertainty for jumping from a state e_i, using action a_k:

$$H(e_i, a_k) = - \sum_{j \in \{1,\ldots,n\}, j \neq i} p_{i,k,j} log(p_{i,k,j})$$

We construct H_1 by summing all the $H(e_i, a_k)$ associated to each state e_i:

$$H_1 = \frac{1}{n.q} \sum_{i \in \{1,\ldots,n\}, k \in \{1,\ldots,q\}} H(e_i, a_k) \qquad (1)$$

A second term, called $H(e_i, e_j)$, characterizes the uncertainty on the action utilized, given that the state e_i was produced at time t and the state at time t+1 was e_j:

$$H(e_i, e_j) = - \sum_{k \in \{1,\ldots,q\}, j \neq i} p_{i,k,j} log(p_{i,k,j})$$

We construct H_2 by summing all the $H(e_i, e_j)$ associated to each couple of states (e_i, e_j):

$$H_2 = \frac{1}{n(n-1)} \sum_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,n\}, i \neq j} H(e_i, e_j) \qquad (2)$$

H_1 is usually less difficult to minimize than H_2 because a_k which is fixed for H_1 is completely controlled by the system: the decision of executing a_k leads identical physical executions. However, this may not be true: imagine that a robot has decided to go 10 cm straight; in reality, the physical property of its environment could force it to go less that 10 cm (because there is an obstacle, for example).

It is well-known that $H(e_i, a_k)$ is minimum and equals 0 if and only if a unique $p_{i,k,j}$ is non zero and equals 1, when i and k are fixed. In the same way, $H(e_i, e_j)$ is minimum and equals 0 if and only if a unique $p_{i,k,j}$ is non zero and equals 1, when i and j are fixed. This specifies the Ideal Context case.

In practice, H_1 and H_2 are computed, when the $p_{i,k,j}$ are known. A very simple algorithm for computing the transition probabilities is as follows:

Produce a sufficiently wide quantity of raw input data and, for each input, deduce the state e_i (using the SRP), apply an action a_k chosen randomly, get the next raw input data after the execution of a_k and notice the state e_j.

Hence, H_1 and H_2 may be deduced outside the learning process.

2.3 Relation between CQ measures and the SRP performance

We test the efficiency of H_1 and H_2 in our experimental environment (see 1.3), without utilizing the RLT. By *efficiency*, we mean:

- H_1 and H_2 must be monotonic functions of the noise amplitude
- The variations of H_1 and H_2 must be sufficiently high when the SRP performance varies.

We use the procedure described in 1.2 to compute H_1 and H_2, for the three types of noises GN, OLN and RSN (see 1.3), with varying values of σ, r_o and r_r. The results are displayed in the figure 1.

A first observation permits to notice that H_1 and H_2 are far from 0 when the amplitude of noise is 0 (graphs (a) and (b)). Thus the context of the RLT is far from being ideal, referring to our CQ measures, even if there is no noise: this is due to the design of the states in itself.

H_1 appears to be clearly better than H_2: for the GN and OLN noises, the amplitude of H_2 is low whereas there is a jump near $r_r = 0$ for the RSN noise and a low variation for $r_r >> 0$. The graph (c) (log/log scale) shows a linear relation between H_1 and the three amplitudes of noise, when they are small enough. This means that H_1 may be modeled by the relation $H_1 = b.\tau^a$ where τ is the slope of the lines in the graph (c).

The behavior of H_2 is surprising at first: there is little variation when σ and r_o vary considerably. We have to remember that we only add noise to θ. There is a very interesting consequence of this fact: given a state e_i and an action a_k, θ is not a discriminant variable for predicting the next state e_j. In fact, this is a consequence of the nature of the benchmark. The variable that is the most important for a state change is $\dot{\theta}$ because it varies very fastlty.

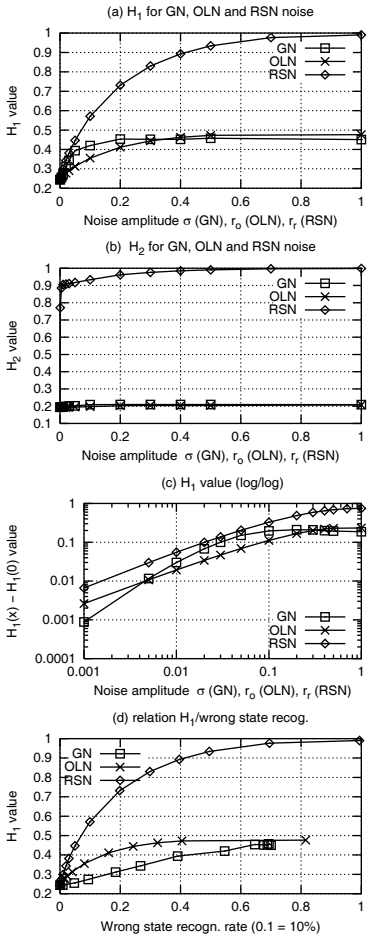

Figure 1. Relation between H_1 and the SRP performance

Finally, the graph (d) shows that H_1 is a monotonic function of the SRP performance (the rate of good state recognition), for the three types of noises. The variations are quite regular when SRP performance varies. In the following we will keep H_1 as the unique CQ measure.

3 Relation between the learning performance and the Context Quality measure

3.1 Model of the learning performance

The learning performance, for a learning trial, is the number of consecutive steps in which the cart/pole is balanced. In our case, the maximum number is fixed to 100 million steps (see 1.3). Hence, the learning performance over the 2000 trials can be modeled by a random variable N, which represents the consecutive steps, and the probability $P(N = n)$. Our first goal is to specify the nature of P.

The graph (a) of the figure 2 illustrates the repartition

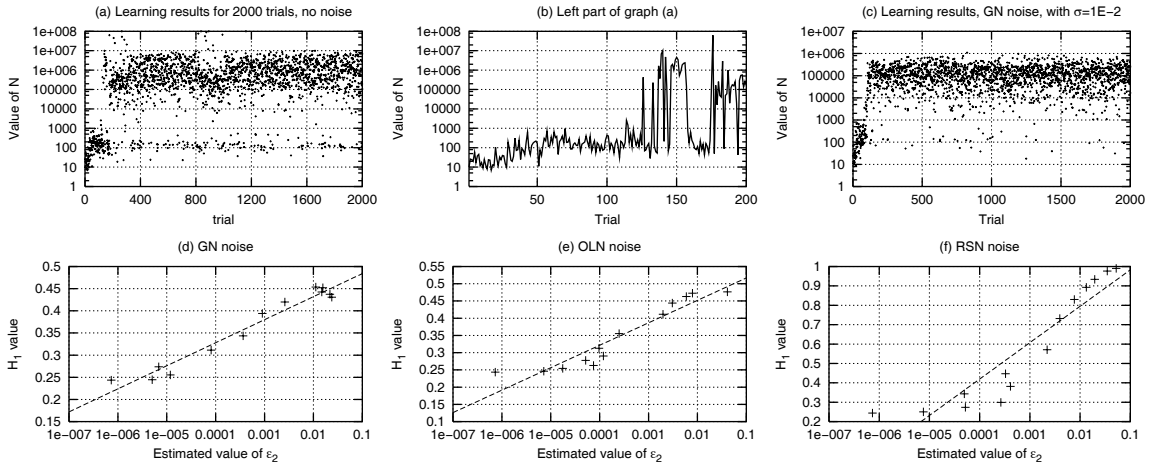

Figure 2. Relation between CQ measure and the frequency of failure (type 2) during the learning phase

of N over the 2000 trials, for a context in which no noise has been added. The classical learning phase consists of the 200 first trials (see graph (b)). After those trials leading once to almost a success (in the trial 175, the cart/pole was balanced for about 50 million steps), we can see that there is not real improvement of the action policy of the system. The values of N seem to be parted into 2 sub-bands of values (one is $[10^5, 10^7]$ and one is around 100). This observation is confirmed if a Gaussian noise is added (graph (c)).

Our goal is not to discuss the effectiveness of the RLT and the context of the RLT we have chosen: we want to explain those results regarding to the H_1 measure (see 2.2). In a first step, we produce a statistical model corresponding to the last results. This model is based on the fact that there exists two kind of independent causes that explain the failure in a trial. The system may jump to the failure state randomly with a small probability ϵ_1 (error of type 1) and ϵ_2 (error of type 2):

$$P(N = n) = p\epsilon_1(1-\epsilon_1)^n + (1-p)\epsilon_2(1-\epsilon_2)^n$$

Where $p \in [0, 1[$.

Following this equation, $E[N]$ and $Var[N]$ may be calculated:

$$E[N] = \frac{p}{\epsilon_1} + \frac{1-p}{\epsilon_2}$$
$$Var[N] = p\frac{p+2-\epsilon_1}{{\epsilon_1}^2} + (1-p)\frac{3-p-\epsilon_2}{{\epsilon_2}^2} + \frac{p(1-p)}{\epsilon_1\epsilon_2} \quad (3)$$

In practice, we compute $E[N]$ and $Var[N]$ from the experimental data. But there are three unknown parameters. In our case, the sub-bands are clearly separated so that ϵ_1 and ϵ_2 are far from each other. So, the parameter p may be estimated independently by counting the number of occurrences of N that are less than 1000. The value of ϵ_1 and ϵ_2 will be deduced by using the former equation.

3.2 Model of the influence of the context on the learning performances

Experiments which are not included in this document have shown that ϵ_1 (associated with a value of N less than 1000) is independent of the nature and the amplitude of noise. The cause of the error in this case might be probably attributed to a "bad" initial value for $(x, \dot{x}, \theta, \dot{\theta})$. The initialization of the system is clearly one of the context components, but it is not taken into account by our CQ measure H_1.

In the paragraph 2.3, we have shown the relation between ESP and H_1. We have just given a model of ESP (equation 3) in which p may be easily estimated and ϵ_1 is a constant when the amplitude of noise varies. From data, we found $\epsilon_1 = 5.7110^{-3}$. What about ϵ_2 ? Is the second source of failures (associated with ϵ_2 in the equation 3) correlated with H_1 ? The graphs (d),(e) and (f) of the figure 2 give a clearly positive answer. Moreover, the relation between ϵ_2 and H_1 may be modeled with the following equation:

$$H_1 = a.log(\epsilon_2) + b \quad (4)$$

It is interesting to notice that the estimated values for a and b are similar for GN and OLN: a=0.023,b=0.54 for GN whereas a=0.028,b=0.58 for OLN. For RSN, the values are quite different: a=0.082,b=1.17 .

But ϵ_2 is not only impacted by the amplitude of noise. Even if there is no noise, the relation 4 is applicable. That means the error source associated to ϵ_2 do not include exclusively the noise, but also probably the design of the states itself.

The relation 4 is very strong because, when H_1 is known (before the learning process), there is a possibility to give the distribution N. Hence, it is possible to predict statistically the performances of RLT.

4 Conclusion and perspective

We have postulated that the context of a learning algorithm is as crucial as the algorithm itself. This article aims to quantify the contextual parameters influence on the performance of a reinforcement learning technique. Our work focuses on the case of the state recognition process which input is the raw data gathered by the system and the output is a state in which the system is supposed to be. This process is clearly contextual and have a high influence on the quality of the results when the raw input data are noisy.

Our experiments are based on the Cart Pole Balancing benchmark. In this case, we prove (section 2) that the Shannon entropy may be utilized to quantify the degradation of the context quality when three types of noise with different amplitudes are applied on the raw input data. We also show (par. 3.2) that, even if there is no noise, the design of states may be a source of failure, which can be partially predicted by looking at the value of the Context Quality measure. For having these results, we build a statistical model of the distribution of the Cart Pole Balancing performance over the learning trials (par. 3.1). Lastly, we express a relation between the Context Quality measure and the recognition process performance (par. 3.2).

What about the generality of our context quality measure ? Undoubtedly, there exists limitations: some contextual parameters do not influence the measure, but have an impact on the performance of the learning algorithm. In particular, the parameters involved in the decision process (mixing exploration and exploitation) are of high importance but are not taken into account. The specification of our measure limits ourself to the influence of the state recognition process. For pure optimization problems, this process is not submitted to uncertainty. The real interest lies on the problems in which a state is difficult to build a priori: this is the case in mobile robotics, even if the noise is low, because we do not always have a model of the mapping between the sensors values and the important structures of the environment.

An ongoing work is carried out to incrementally build the internal states of the robot in order to minimize our quality context measure. Some pieces of work have shown that states which take into account data over time are associated to a better quality, even if the noise is low. The results obtained on the Cart Pole Balancing problem suggest that the inertia of a dynamic system might impact badly the context quality, hence the learning performance. Taking into account data over time is probably a manner of reducing this cause.

References

[1] A.G. Barto, R.S. Sutton, and C.W. Anderson. Neurolike adaptive elements that can solve difficult learning control problems. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC13, 1983, 834–846.

[2] B. Bouchon. Entropic models: a general framework for measures of uncertainty and information. *Logic in Knowledge-Based Systems, Decision and Control*, 1988, 93–105.

[3] F. Davesne. *Etude de l'émergence de facultés d'apprentissage fiables et prédictibles d'actions réflexes, à partir de modèles paramétriques soumis à des contraintes internes*. PhD thesis, University of Evry, France, 2002.

[4] F. Davesne and C. Barret. Constraint based memory units for reactive navigation learning. In *European Workshop on Learning Robots*, Lausanne. 1999.

[5] F. Davesne and C. Barret. Reactive navigation of a mobile robot using a hierarchical set of learning agents. In *IROS'99*, Kyongyu, Korea. 1999.

[6] P. Dayan and T.J. Sejnowski. Td(λ) converges with probability 1. *Machine Learning*, 14, 1994, 295–301.

[7] P.Y. Glorennec. Fuzzy q-learning and dynamical fuzzy q-learning. In *Proc. of the 3th IEEE Fuzzy systems conference*, Orlando. 1994.

[8] S. Harnad. Cognition and the symbol grounding problem. *Electronic symposium on computation*, 1992.

[9] M.L. Littman, A. Cassandra, and L. Kaelbling. Learning policies for partially observable environments: Scaling up. In Armand Prieditis and Stuart Russell (Eds.), *Twelfth Int. Conf. on Machine Learning*, San Francisco, CA, USA. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995, 362–370.

[10] M. Mataric. Integration of representation into goal-driven behavior-based robots. *IEEE Trans. Robotics and Automation*, 8(3), 1992.

[11] R. Munos. Variable resolution discretization for high-accuracy solutions of optimal control problem. *Int. Joint Conf. on Artificial Intelligence*, 1999.

[12] M.D. Pendrith. Reinforcement learning in situated agents: Some theoretical problems and practical solutions. In *8th European Workshop on Learning Robots*, Lausanne. 1999.

[13] M.D. Pendrith and M.J. McGarity. An analysis of direct reinforcement learning in non-markovian domains. *The Fifteenth International Conference on Machine Learning*, 1998.

[14] J. Peng and R.J. Williams. Incremental multi-step q-learning. *Machine Learning*, 22, 1996, 283–290.

[15] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An introduction*. MIT Presss, Cambridge, MA, 1998.