Author prepared draft.
D. Vincze, Sz. Kovács: Fuzzy Rule Interpolation-based Q-learning
CIIS2009, first workshop on Computational Intelligence in Information Science, Miskolc, Hungary, May 25-26, 2009 as part of:
SACI 2009 5th International Symposium on Applied Computational Intelligence and Informatics May 28-29, 2009, Timisoara, Romania, pp. 55-59.

Fuzzy Rule Interpolation-based Q-learning

Dávid Vincze, Szilveszter Kovács
Department of Information Technology, University of Miskolc, Miskolc, Hungary
david.vincze@iit.uni-miskolc.hu
szkovacs@iit.uni-miskolc.hu

Abstract—**Reinforcement learning is a well known topic in computational intelligence. It can be used to solve control problems in unknown environments without defining an exact method on how to solve problems in various situations. Instead the goal is defined and all the actions done in the different states are given feedback, called reward or punishment (positive or negative reward). Based on these rewards the system can learn which action is considered the best in a given state. A method called Q-learning can be used for building up the state-action-value function. This method uses discrete states. With the application of fuzzy reasoning the method can be extended to be used in continuous environment, called Fuzzy Q-learning (FQ-Learning). Traditional Fuzzy Q-learning uses 0-order Takagi-Sugeno fuzzy inference. The main goal of this paper is to introduce Fuzzy Rule Interpolation (FRI), namely the FIVE (Fuzzy rule Interpolation based on Vague Environment) to be the model applied with Q-learning (FRIQ-learning). The paper also includes an application example: the well known cart pole (reversed pendulum) problem is used for demonstrating the applicability of the FIVE model in Q-learning.**

I. Introduction

The strength of reinforcement learning lies in the fact that it does not specify how to solve a particular problem, instead the final goal is defined and the state-action-value function is learned based on rewards or punishments (negative rewards) given by the environment on the goodness of the selected action in the given observable state. So it focuses on what to do not how to do. Reinforcement learning techniques are a kind of trial-and-error style techniques adapting to dynamic environment via incremental iterations. The primary ideas of reinforcement learning techniques (dynamical system state and the idea of "optimal return" / "value" function) are inherited from optimal control and dynamic programming [3]. Finding an optimal policy by building the state-value- or action-value-function is a common goal of the reinforcement learning strategies [19]. The state-value-function $V^{\pi}(s)$, is a function of the expected return (a function of the cumulative reinforcements), related to a given state $s \in S$ as a starting point, following a given π policy. These rewards (reinforcements) are the expression of the desired goal of the learning agent as a kind of evaluation following the previous action (in spite of the instructive manner of error feedback based approximation techniques, like the gradient descent optimisation). The policy is the description of the agent behaviour, in the form of mapping between the agent states and the corresponding suitable actions. The action-value function $Q^{\pi}(s,a)$ is a function of the expected return, in case of taking action $a \in A_s$ in a given state s, and then following a given policy π. Having the action-value-function, the optimal (greedy) policy, which always takes the optimal (the greatest estimated value) action in every state, can be constructed as [19]:

$$\pi(s) = \arg\max_{a \in A_s} Q^{\pi}(s,a) \qquad \textbf{(1)}$$

Namely for estimating the optimal policy, the action-value function $Q^{\pi}(s,a)$ is needed to be approximated. Having a complex task to adapt, both the number of possible states and the number of the possible actions could be an extremely high value. This implies the problem that it takes a considerable amount of computing resources and computation time, which is a significant drawback of reinforcement learning; however there are some cases where a distributed approach with continuous reward functions can reduce these resource needs [15]. Generally reinforcement learning methods can lead to results in practically acceptable time only in relatively small state and action spaces.

With the introduction of fuzzy models, the discrete Q-learning can be extended to continuous state and action space, which in case of suitably chosen states can lead to the reduction of the size of the state-action space [12].

A. Q-learning and Fuzzy Q-learning

The purpose of the Q-learning is finding the fixed-point solution Q of the Bellman Equation [3] through iteration. In discrete environment *Q-Learning* [22], the action-value-function is approximated by the following iteration:

$$Q_{i,u} \approx \widetilde{Q}_{i,u}^{k+1} = \widetilde{Q}_{i,u}^{k} + \Delta\widetilde{Q}_{i,u}^{k+1} = \widetilde{Q}_{i,u}^{k} + \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v \in U} \widetilde{Q}_{j,v}^{k+1} - \widetilde{Q}_{i,u}^{k} \right) \qquad \textbf{(2)}$$

$\forall i \in I, \forall u \in U$, where $\widetilde{Q}_{i,u}^{k+1}$ is the *k+1* iteration of the action-value taking the u^{th} action A_u in the i^{th} state S_i, S_j is the new (j^{th}) observed state, $g_{i,u,j}$ is the observed reward completing the $S_i \rightarrow S_j$ state-transition, γ is the discount factor and $\alpha_{i,u}^{k} \in [0,1]$ is the step size parameter (which can change during the iteration steps), *I* is the set of the discrete possible states and *U* is the set of the discrete possible actions. Many solutions exist [1], [4], [5], [6] for applying this iteration to continuous environment by adopting fuzzy inference (Fuzzy Q-Learning). Traditionally the simplest FQ-Learning, the 0-order Takagi-Sugeno Fuzzy Inference model is the most common. Therefore in this paper this one is studied (a slightly modified, simplified version of the Fuzzy Q-Learning introduced in [1] and [6]). In this case, for characterizing the value function *Q(s,a)* in continuous state-action space, the order-0 Takagi-Sugeno Fuzzy

Inference System approximation $\widetilde{Q}(s,a)$ is adapted in the following manner:

$$\textbf{If } s \textbf{ is } S_i \textbf{ And } a \textbf{ is } A_u \textbf{ Then } \widetilde{Q}(s,a)=Q_{i,u}, \quad i\in I, u\in U, \tag{3}$$

where S_i is the label of the i^{th} membership function of the n dimensional state space, A_u is the label of the u^{th} membership function of the one dimensional action space, $Q_{i,u}$ is the singleton conclusion and $\widetilde{Q}(s,a)$ is the approximated continuous state-action-value function. Having the approximated state-action-value function $\widetilde{Q}(s,a)$, the optimal policy can be constructed by function (1). Setting up the antecedent fuzzy partitions to be *Ruspini partitions*, the order-0 Takagi-Sugeno fuzzy inference forms the following approximation function:

$$\widetilde{Q}(s,a)=\sum_{i_1,i_2,\cdots,i_N,u}^{I_1,I_2,\ldots,I_N,U}\prod_{n=1}^{N}\mu_{i_n,n}(s_n)\cdot\mu_u(a)\cdot q_{i_1 i_2 \ldots i_N u} \tag{4}$$

where $\widetilde{Q}(s,a)$ is the approximated state-action-value function, $\mu_{i_n,n}(s_n)$ is the membership value of the i_n^{th} state antecedent fuzzy set at the n^{th} dimension of the N dimensional state antecedent universe at the state observation s_n, $\mu_u(a)$ is the membership value of the u^{th} action antecedent fuzzy set of the one dimensional action antecedent universe at the action selection a, $q_{i_1 i_2 \ldots i_N u}$ is the value of the singleton conclusion of the $i_1,i_2,\ldots,i_N,u^{\text{-th}}$ fuzzy rule. Applying the approximation formula of the Q-learning (2) for adjusting the singleton conclusions in (4), leads to the following function:

$$q^{k+1}_{i_1 i_2 \ldots i_N u}=q^{k}_{i_1 i_2 \ldots i_N u}+\prod_{n=1}^{N}\mu_{i_n,n}(s_n)\cdot\mu_u(a)\cdot\Delta\widetilde{Q}^{k+1}_{i,u}=$$
$$=q^{k}_{i_1 i_2 \ldots i_N u}+\prod_{n=1}^{N}\mu_{i_n,n}(s_n)\cdot\mu_u(a)\cdot\alpha^{k}_{i,u}\cdot\left(g_{i,u,j}+\gamma\cdot\max_{v\in U}\widetilde{Q}^{k+1}_{j,v}-\widetilde{Q}^{k}_{i,u}\right) \tag{5}$$

where $q^{k+1}_{i_1 i_2 \ldots i_N u}$ is the *k+1* iteration of the singleton conclusion of the $i_1 i_2 \ldots i_N u^{\text{th}}$ fuzzy rule taking action A_u in state S_i, S_j is the new observed state, $g_{i,u,j}$ is the observed reward completing the $S_i \rightarrow S_j$ state-transition, γ is the discount factor and $\alpha^{k}_{i,u}\in[0,1]$ is the step size parameter. The $\mu_{i_n,n}(s_n)\cdot\mu_u(a)$ is the partial derivative of the conclusion of the 0-order Takagi-Sugeno fuzzy inference $\widetilde{Q}(s,a)$ with respect to the fuzzy rule consequents $q_{u,i}$ according to (4), required for the applied steepest-descent optimization method. The $\widetilde{Q}^{k+1}_{j,v}$ and $\widetilde{Q}^{k}_{i,u}$ action-values can be approximated by equation (4).

II. FRI-based Fuzzy Q-learning

In the followings the proposed FIVE FRI based Q-learning method will be introduced in more details.

A. The FIVE FRI method

Various FRI techniques exist, a comprehensive overview of the recent existing FRI methods can be found in [2]. FIVE is one of these techniques, it is an application oriented fuzzy rule interpolation method (introduced in [11], [9] and [13]), it is fast and serves crisp conclusions directly so there is no need for an additional defuzzification step. Also FIVE has been already proved to be capable of serving the requirements of practical applications [21].

The main idea of the FIVE is based on the fact that most of the control applications serves crisp observations and requires crisp conclusions from the controller. Adopting the idea of the vague environment (VE) [8], FIVE can handle the antecedent and consequent fuzzy partitions of the fuzzy rule base by scaling functions [8] and therefore turn the fuzzy interpolation to crisp interpolation. The idea of a VE is based on the similarity (in other words: indistinguishability) of the considered elements. In VE the fuzzy membership function $\mu_A(x)$ is indicating level of similarity of x to a specific element a that is a representative or prototypical element of the fuzzy set $\mu_A(x)$, or, equivalently, as the degree to which x is indistinguishable from a [8]. Therefore the α-cuts of the fuzzy set $\mu_A(x)$ are the sets which contain the elements that are (1-α)-indistinguishable from *a*. Two values in a VE are ε-distinguishable if their distance is greater than ε. The distances in a VE are weighted distances. The weighting factor or function is called scaling function (factor) [8]. If a VE of a fuzzy partition (the scaling function or at least the approximate scaling function [11], [13]) exists, the member sets of the fuzzy partition can be characterized by points in that VE (see e.g. scaling function *s* on Fig. 1). Therefore any crisp interpolation, extrapolation, or regression method can be adapted very simply for FRI [11], [13]. Because of its simple multidimensional applicability, in FIVE the Shepard operator based interpolation (first introduced in [17]) is adapted (see e.g. Fig. 1).

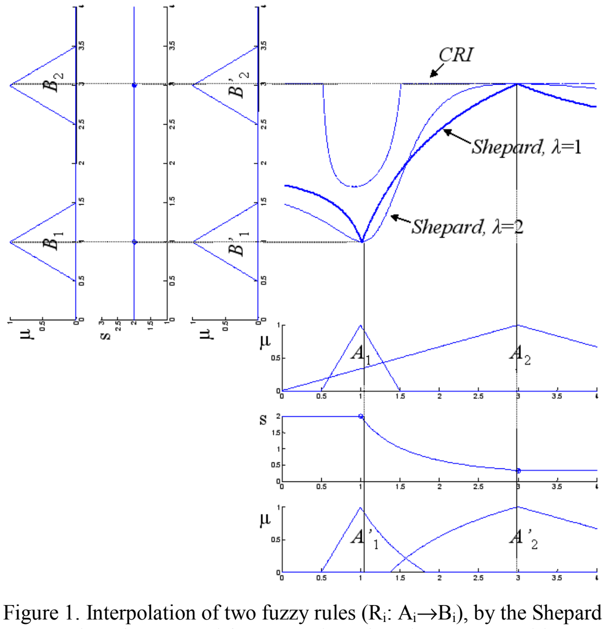

Figure 1. Interpolation of two fuzzy rules (R_i: A_i→B_i), by the Shepard operator based *FIVE*, and for comparison the min-max *CRI* with COG defuzzification.

The code of the FIVE FRI along with other FRI methods is freely available as a MatLab FRI Toolbox [7], and they can be downloaded from [24] and [25].

B. The FIVE FRI-based Q-learning

Replacing the zero-order Takagi-Sugeno fuzzy model of the FQ-learning with the FIVE model leads to the proposed FRIQ-learning method.

Introducing the FIVE FRI in Q-learning gives the possibility of omitting rules (action-state values) from the rule base and gaining the potentiality of applying the proposed method in higher state dimensions with a reduced rule-base sized action-state space. An example for effective rule base reduction by FRI FIVE in a given situation is introduced in [18].

The FIVE FRI based fuzzy model in case of singleton rule consequents [10] can be expressed by the following formula:

$$\tilde{Q}(s,a)=\begin{cases} q_{i_1 i_2 .. i_N u} & \text{if } \mathbf{x}=\mathbf{a}_k \text{ for some } k, \\ \left(\sum_{k=1}^{r} q_{i_1 i_2 .. i_N u} \big/ \delta_{s,k}^{\lambda}\right) \Big/ \left(\sum_{k=1}^{r} 1 \big/ \delta_{s,k}^{\lambda}\right) & \text{otherwise.} \end{cases} \tag{6}$$

where the fuzzy rules R_k have the form:

If $\mathrm{x}_1 = A_{\mathrm{k},1}$ **And** $\mathrm{x}_2 = A_{\mathrm{k},2}$ **And** … **And** $\mathrm{x}_m = A_{\mathrm{k},m}$ **Then** $\mathrm{y} = \mathrm{c}_\mathrm{k}$, **(7)**

$\delta_{s,k}$ is the scaled distance:

$$\delta_{s,k} = \delta_s(\mathbf{a}_k,\mathbf{x}) = \left[\sum_{i=1}^{m}\left(\int_{\mathrm{a}_{k,i}}^{\mathrm{x}_i} s_{\mathrm{X}_i}(\mathrm{x}_i)\,d\mathrm{x}_i\right)^2\right]^{1/2}, \tag{8}$$

and s_{X_i} is the i^{th} scaling function of the m dimensional antecedent universe, $\mathbf{x}$ is the m dimensional crisp observation and $\mathbf{a}_k$ are the cores of the m dimensional fuzzy rule antecedents $\boldsymbol{A}_k$.

Applying the FIVE FRI method with singleton rule consequents (6) to be the model of the state-action-value function, we get:

$$\tilde{Q}(s,a)=\begin{cases} q_{i_1 i_2 .. i_N u} & \text{if } \mathbf{x}=\mathbf{a}_k \text{ forsome} k, \\ \sum_{i_1,i_2,\cdots;i_N,u}^{I_1,I_2,\ldots,I_N,U} \prod_{n=1}^{N}\left(1/\delta_{s,k}^{\lambda}\right) \Big/ \left(\sum_{k=1}^{r} 1/\delta_{s,k}^{\lambda}\right) \cdot q_{i_1 i_2 .. i_N u} & \text{otherwise} \end{cases} \tag{9}$$

where $\tilde{Q}(s,a)$ is the approximated state-action-value function.

The partial derivative of the model consequent $\tilde{Q}(s,a)$ with respect to the fuzzy rule consequents $q_{u,i}$, required for the applied fuzzy Q-learning method (5) in case of the FIVE FRI model from (9) can be expressed by the following formula (according to [14]):

$$\frac{\partial \tilde{Q}(s,a)}{\partial q_{i_1 i_2 .. i_N u}} = \begin{cases} 1 & \text{if } x = a_k \text{ for some } k, \\ \left(1/\delta_{s,k}^{\lambda}\right) \Big/ \left(\sum_{k=1}^{r} 1/\delta_{s,k}^{\lambda}\right) & \text{otherwise.} \end{cases} \tag{10}$$

where $q_{u,i}$ is the constant rule consequent of the k^{th} fuzzy rule, $\delta_{s,k}$ is the scaled distance in the vague environment of the observation, and the k^{th} fuzzy rule antecedent, λ is a parameter of Shepard interpolation (in case of the stable multidimensional extension of the Shepard interpolation it equals to the number of antecedents according to [20]), x is the actual observation, r is the number of the rules.

Replacing the partial derivative of the conclusion of the 0-order Takagi-Sugeno fuzzy inference (5) with the partial derivative of the conclusion of FIVE (10) with respect to the fuzzy rule consequents $q_{u,i}$ leads to the following equation for the Q-Learning action-value-function iteration:

if $\mathbf{x}=\mathbf{a}_k$ for some k:

$$q_{i_1 i_2 .. i_N u}^{k+1} = q_{i_1 i_2 .. i_N u}^{k} + \Delta\tilde{Q}_{i,u}^{k+1} = = q_{i_1 i_2 .. i_N u}^{k} + \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v\in U} \tilde{Q}_{j,v}^{k+1} - \tilde{Q}_{i,u}^{k}\right) \tag{11}$$

otherwise

$$q_{i_1 i_2 .. i_N u}^{k+1} = q_{i_1 i_2 .. i_N u}^{k} + \prod_{n=1}^{N}\left(1/\delta_{s,k}^{\lambda}\right) \Big/ \left(\sum_{k=1}^{r} 1/\delta_{s,k}^{\lambda}\right) \cdot \Delta\tilde{Q}_{i,u}^{k+1} =$$

$$= q_{i_1 i_2 .. i_N u}^{k} + \prod_{n=1}^{N}\left(1/\delta_{s,k}^{\lambda}\right) \Big/ \left(\sum_{k=1}^{r} 1/\delta_{s,k}^{\lambda}\right) \cdot \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v\in U} \tilde{Q}_{j,v}^{k+1} - \tilde{Q}_{i,u}^{k}\right)$$

where $q_{i_1 i_2 \ldots i_N u}^{k+1}$ is the *k+1* iteration of the singleton conclusion of the $i_1 i_2 \ldots i_N u$ th fuzzy rule taking action A_u in state S_i, S_j is the new observed state, $g_{i,u,j}$ is the observed reward completing the $S_i \to S_j$ state-transition, γ is the discount factor and $\alpha_{i,u}^{k} \in [0,1]$ is the step size parameter.

As in the previous chapter the $\tilde{Q}_{j,v}^{k+1}$ and $\tilde{Q}_{i,u}^{k}$ action-values can be approximated by equation (11), which now uses the FIVE FRI model.

III. APPLICATION EXAMPLE

As a benchmark for the proposed FRI based Q-learning method, the well known cart-pole (reversed pendulum) problem is chosen as an application example in this paper. (An implementation by José Antonio Martin H. which uses SARSA [16] (a Q-learning method) in discrete space is freely available from [26].) In order to make the comparison easier, the application example introduced in [26] was extended to adapt the FIVE FRI model.

For easier comparability purposes in the application example, the discrete Q-learning, and the proposed FRIQ-learning had the same state-action space resolution. In the discrete case the resolution means the number of the discrete cases, in the fuzzy model case, these are the cores of the fuzzy sets in the antecedent fuzzy partitions.

The example program runs through episodes, where an episode means a cart-pole simulation run. An episode is considered to be successfully finished (positive reinforcement) if the number of iterations (steps) reaches one thousand while the pole stays up without the cart

TABLE I.
SAMPLE RULES FROM THE Q "CALCULATION" RULE BASE

R#	s_1	s_2	s_3	s_4	a	q
0001	N	N	N12	N	AN10	0
0512	N	Z	N3	N	AN3	0
1024	N	P	P6	N	AP5	0
2048	P	P	N3	P	Z	0
2268	P	P	P12	P	AP10	0

TABLE II.
RULES WHICH HAVE WEIGHTS GREATER THAN 0.01 IN CASE OF A SPECIFIED OBSERVATION

R#	s_1	s_2	s_3	s_4	a	w
0013	N	N	N12	N	AP2	0.020319
0054	N	N	N9	N	AP1	0.026501
0055	**N**	**N**	**N9**	**N**	**AP2**	**0.162300**
0056	N	N	N9	N	AP3	0.026501
0097	N	N	N6	N	AP2	0.020319
0432	N	Z	N9	N	AP1	0.010130
0433	N	Z	N9	N	AP2	0.029851
0434	N	Z	N9	N	AP3	0.010130
1147	P	P	N12	N	AP2	0.021033
1188	P	P	N9	N	AP1	0.027540
1189	**P**	**P**	**N9**	**N**	**AP2**	**0.175810**
1190	P	P	N9	N	AP3	0.027540
1231	P	P	N6	N	AP2	0.021033
1566	P	Z	N9	N	AP1	0.010398
1567	P	Z	N9	N	AP2	0.031080
1568	P	Z	N9	N	AP3	0.010398

crashing into the walls. Otherwise the episode is considered to be failed (negative reinforcement).

Some sample rules from the rule base used for calculating the q values can be seen in Table I. The rule antecedent variables are the following: s_1 – shift of the pendulum, s_2 – velocity of the pendulum, s_3 – angular offset of the pole, s_4 – angular velocity of the pole, a – compensation action of the cart. The linguistic terms used in the antecedent parts of the rules are: Negative (N), Zero (Z), Positive (P), the multiples of three degrees in [-12,12] degree interval (N12, N9, N6, N3, Z, P3, P6, P9, P12) and for the actions: from negative to positive in one tenth steps (AN10-AP10, Z). For the easier comparison purposes, the above resolution of the state-action spaces are the same as the resolutions of the discrete Q-learning implementation selected as the reference example (introduced in [26]). The consequents (q) are initialized with zero values. These values will be then updated while learning according to (11). The fuzzy rules are defined in the following form:

R_i: **(12)**

If $s_1 = A_{1,i}$ **and** $s_2 = A_{2,i}$ **and** $s_3 = A_{3,i}$ **and** $s_4 = A_{4,i}$ **and** $a = A_{5,i}$ **Then** $q = B_i$

For testing and validation purposes the randomized action selection (see [26]) (state-action space exploration) was disabled temporarily. With using the same state descriptors with both the original [26] and FRIQ-learning version the results were exactly the same as expected (the state variables hit exactly the cores of the fuzzy sets at the antecedent side, hence the value of the derivative - weight of the rule - was equal to 1).

In continuous environment there are so many states that exploring the whole universe can require tremendous amount of computing power and time. In order to get a useable result in acceptable time but to use the extended capabilities of FRI based Q-learning, one more state value was added to the states [26]. The shift values (s_1) which were originally -1 and 1 are now extended with a new, zero value: -1, 0, 1. The rest of the rule base remained the same as with the two states, hence handling the new 0 value in s_1 is the job of the FRI based fuzzy inference model. The results of the simulation with the original and the extended s_1 values can be seen on Fig 2. and Fig 3. Fig 2. shows how many iteration steps could the cart survives both with the original two s_1 state values (red) and the extended three s_1 values (green). Fig 3. shows the cumulated rewards in each iteration cycle (red: two s_1 states, green: three s_1 states). The figures show that both versions learn a suitable state-action-value function, and the one with three s_1 states gives better results at first, but converges slower. Table II. shows the weights of rule consequent (q) updates in case of a specified observation with the new s_1 state: s_1 = Z, s_2 = N, s_3 = N9, s_4 = N, a = AP2. There are a lot of rules with so small weights that they do not affect the q values considerably, hence only the rules with weights more than 0.01 are shown on Table II. Except s_1 all of the antecedents hit exactly the cores of the fuzzy sets. The two heaviest rules are typesetted bold, and it can be clearly seen that with using FIVE the rule updates are positive and negative with nearly the same frequencies (requiring the missing zero value).

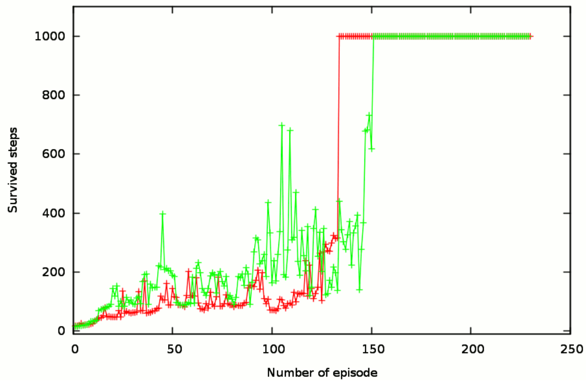

Figure 2. Steps survived with the original two states (red) and with the extended three states (green).

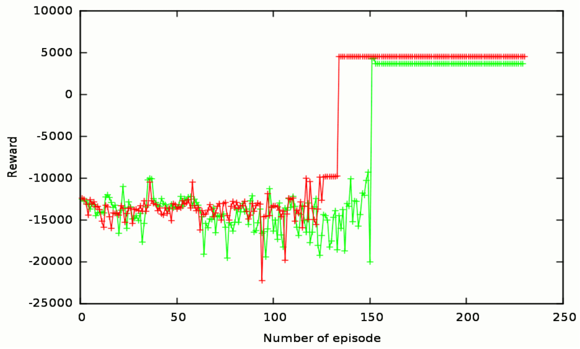

Figure 3. The cumulative rewards the original two states (red) and with the extended three states (green).

The source code of the FIVE extended cart-pole simulation program can be freely accessed at [23].

CONCLUSIONS

With the introduction of FIVE FRI in Q-learning instead of discrete state-action spaces, continuous spaces could be applied. Applying continuous spaces can lead to better resolution providing more precise description of the state-action pairs. The application example and the numerical results prove that the proposed FRI model based method performs similarly to the discrete solution in case of the same action-state resolution and circumstances. Even if in this case the FRI based Q-learning performs in a similar way to the discrete solution, it holds the potentialities of action-state space (i.e. rule base in FRI case) reduction. The way of rule base reduction with keeping the performance of the original system is beyond the scope of this paper. It needs the identification of the insignificant rules to be dynamically omitted from the rule base, making it smaller, exploiting the real advantages of the proposed FIVE based FRIQ-learning method.

ACKNOWLEDGMENT

This research was partly supported by the Hungarian National Scientific Research Fund grant no: OTKA K77809 and the Intelligent Integrated Systems Japanese Hungarian Laboratory.

REFERENCES

[1] Appl, M.: Model-based Reinforcement Learning in Continuous Environments. Ph.D. thesis, Technical University of München, München, Germany, dissertation.de, Verlag im Internet (2000)

[2] P. Baranyi, L. T. Kóczy, and Gedeon, T. D., "A Generalized Concept for Fuzzy Rule Interpolation", IEEE Trans. on Fuzzy Systems, vol. 12, No. 6, 2004, pp 820-837.

[3] Bellman, R. E.: Dynamic Programming. Princeton University Press, Princeton, NJ (1957)

[4] Berenji, H.R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems (1996) pp 2208-2214.

[5] Bonarini, A.: Delayed Reinforcement, Fuzzy Q-Learning and Fuzzy Logic Controllers. In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, (Studies in Fuzziness, 8), Physica-Verlag, Berlin, D, (1996) pp 447-466.

[6] Horiuchi, T., Fujino, A., Katai, O., Sawaragi, T.: Fuzzy Interpolation-Based Q-learning with Continuous States and Actions. Proc. of the 5th IEEE International Conference on Fuzzy Systems, Vol.1 (1996) pp 594-600.

[7] Zs. Cs. Johanyák, D. Tikk, Sz. Kovács, K. W. Wong: Fuzzy Rule Interpolation Matlab Toolbox – FRI Toolbox, Proc. of the IEEE World Congress on Computational Intelligence (WCCI'06), 15th Int. Conf. on Fuzzy Systems (FUZZ-IEEE'06), July 16-21, Vancouver, BC, Canada, Omnipress. ISBN 0-7803-9489-5, 2006, pp. 1427-1433.

[8] F. Klawonn, "Fuzzy Sets and Vague Environments", Fuzzy Sets and Systems, 66, 1994, pp. 207-221.

[9] Sz. Kovács, and L.T. Kóczy, "Approximate Fuzzy Reasoning Based on Interpolation in the Vague Environment of the Fuzzy Rule base as a Practical Alternative of the Classical CRI", Proceedings of the 7th International Fuzzy Systems Association World Congress, Prague, Czech Republic, 1997, 144-149.

[10] Kovács, Sz.: Extending the Fuzzy Rule Interpolation "FIVE" by Fuzzy Observation, Advances in Soft Computing, Computational Intelligence, Theory and Applications, Bernd Reusch (Ed.), Springer Germany, ISBN 3-540-34780-1, pp. 485-497, (2006).

[11] Sz. Kovács, "New Aspects of Interpolative Reasoning", Proceedings of the 6th. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, 1996, pp. 477-482.

[12] Sz. Kovács: SVD Reduction in Continuos Environment Reinforcement Learning, Lecture Notes in Computer Science, Vol. 2206, Computational Intelligence, Theory and Applications, Bernard Reusch (Ed.), Springer, ISBN 3-540-42732-5, pp.719-738, Germany, (2001).

[13] Sz. Kovács, and L.T. Kóczy, "The use of the concept of vague environment in approximate fuzzy reasoning", Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic, vol.12, 1997, pp. 169-181.

[14] Krizsán, Z., Kovács, Sz.: Gradient based parameter optimisation of FRI "FIVE", Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, Hungary, November 6-8, ISBN 978-963-7154-82-9, pp. 531-538, (2008).

[15] José Antonio Martin H., Javier De Lope: A Distributed Reinforcement Learning Architecture for Multi-Link Robots. 4th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2007), 2007

[16] Rummery, G. A., Niranjan, M.: On-line Q-learning using connectionist systems. CUED/F-INFENG/TR 166, Cambridge University, UK. (1994)

[17] D. Shepard, "A two dimensional interpolation function for irregularly spaced data", Proc. 23rd ACM Internat. Conf., 1968, pp. 517-524.

[18] Sz. Kovács: Interpolative Fuzzy Reasoning in Behaviour-based Control, Advances in Soft Computing, Vol. 2, Computational Intelligence, Theory and Applications, Bernd Reusch (Ed.), Springer, Germany, ISBN 3-540-22807-1, pp.159-170, (2005).

[19] Sutton, R. S., Barto, A. G.: Reinforcement Learning: An Introduction, MIT Press, Cambridge (1998)

[20] D. Tikk, I. Joó, L. T. Kóczy, P. Várlaki, B. Moser, and T. D. Gedeon (2002). Stability of interpolative fuzzy KH-controllers. Fuzzy Sets and Systems, (125) 1, 105-119.

[21] Vincze, D., Kovács, Sz.: Using Fuzzy Rule Interpolation-based Automata for Controlling Navigation and Collision Avoidance Behaviour of a Robot, IEEE 6th International Conference on Computational Cybernetics, Stara Lesná, Slovakia, November 27-29, ISBN: 978-1-4244-2875-5, pp. 79-84, (2008).

[22] Watkins, C. J. C. H.: Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, Cambridge, England (1989)

[23] The source code of the FIVE extended cart-pole simulation program can be found at:
http://www.iit.uni-miskolc.hu/~vinczed/

[24] The FRI Toolbox is available at: http://fri.gamf.hu/

[25] Some FRI applications are available at:
http://www.iit.uni-miskolc.hu/~szkovacs

[26] The cart-pole example for discrete space can be found at: http://www.dia.fi.upm.es/~jamartin/download.htm