# EXPERT HEURISTIC TUNING DESIGN FOR THE FRIQ-LEARNING

**Tamás Tompa**
*assistant lecturer, University of Miskolc, Department of Information Technology*
*H-3515 Miskolc, Miskolc-Egyetemváros, Hungary, e-mail: tompa@iit.uni-miskolc.hu*

**Szilveszter Kovács**
*associate professor, University of Miskolc, Department of Information Technology*
*H-3515 Miskolc, Miskolc-Egyetemváros, Hungary, e-mail: szkovacs@iit.uni-miskolc.hu*

***Abstract***

*The conventional reinforcement learning (RL) methods (e.g. Q-learning, SARSA, Fuzzy Q-learning) are searching for the solution starting from an empty initially empty knowledgebase, which is then expanded and filled by the problem related knowledge through iterations incrementally. These traditional RL systems do not have any additional external knowledge about the solution, therefore the learning phase may be a long process. Many methods exist which is able to inject external information into the RL system. This RL area is called heuristically accelerated reinforcement learning. The heuristically accelerated version of the fuzzy rule interpolation based Q-learning (FRIQ-learning) is able to incorporate the external expert knowledge in form of fuzzy rule-base into its knowledgebase. In this FRIQ-learning system the expert knowledge is static, it does not change during the learning phase. In the case if the external knowledge is not entirely correct, it can have a negative influence on the system efficiency (e.g. low convergence rate). Thus a methodology is needed, which is able to optimize (tune) the external knowledge rule-base (Q-function) during the learning phase too. The main goal of this paper is to suggest a method for the FRIQ-learning system which may be suitable for optimizing the injected expert knowledgebase (Q-function) too.*

***Keywords:*** *reinforcement learning, heuristically accelerated reinforcement learning, expert knowledgebase, Q-learning, fuzzy Q-learning*

## 1. Introduction

The reinforcement learning (RL) [13] is one of the popular areas of the computational intelligence. These methods and algorithms are working based on the reinforcements (which can be reward or punishment) given by the environment. The agent due to a properly specified reward function get feedback (as reinforcement) from the environment for the executed action in a given state, which modifies its expectations related to the reactions of the environment, then search for the solution through repetitive experiments. The main benefit of these methods that they can start without any preliminary knowledge about the model describing the problem, and they are able to search for the solution and produce the related knowledgebase based on the goal given in the form reinforcements. The RL agent continuously evolve its behaviour based on the reinforcements, and improves its abilities. Many RL algorithms can be found in the literature. The most common ones are the Q-learning [23], the SARSA [12] and their modificated versions, e.g. the fuzzy model based versions [1] [2] [5]. These traditional RL methods start the learning task with an empty knowledgebase, then expanding it through iterations until a final knowledgebase, which is describing the solution, is achieved. In case if external information of can be injected into the

RL system, the number of the required iteration steps can be decreased, and hence the convergence speed, the efficiency of the learning system can be increased.

The heuristically accelerated reinforcement learning (HARL) methods (e.g. HAQL, HA-Q(λ), HA-SARSA(λ), HA-TD(λ)) [3] [4] are giving a possibility to inject external knowledge (as heuristic) into the learning system. Depending on the applied model, the knowledge representation can be different. It is a Q-table in the conventional Q-learning, and a rule-base in the fuzzy based RL systems. In the HARL system [4] the heuristic information is described by a $H_t(s_t, a_t)$ heuristic function, which is specifying which $a_t$ action preferred in the $s_t$ state at the $t$ time. Another solution to describe the external information is the 'GOAL' (Goal-Oriented Agent Language) [6], which determine the action selection policy by "if then" rules.

In the fuzzy model based reinforcement learning systems the knowledgebase is represented by a fuzzy rule-base, therefore the external knowledge (like heuristic information) should also be formalized through fuzzy rule-base [18]. The Fuzzy Rule Interpolation based Q-learning (FRIQ-learning) [18] [21] is an extension of traditional Q-learning with a fuzzy rule interpolation (FRI) model (e.g. 'FIVE' FRI [10]). In this system the Q-function is represented by a sparse fuzzy rule-base. Therefore the heuristic information, as "expert knowledge", can be also given in the form of sparse fuzzy rule-base [15]. In case if the expert rule-base correctly defined, the convergence speed of the FRIQ-learning system can be significantly improve [14] [15]. Otherwise, having incorrect information in the expert defined knowledgebase, it can also have negative influence for the system performance.

The main goal of this paper is to introduce a method for the expert knowledge incorporated FRIQ-learning, which is also suitable for tuning (optimize) the knowledgebase, even in case if the a priori expert heuristic contains incorrect information about the solution.

## 2. Expert knowledge injected FRIQ-learning

The FRIQ-learning (Fuzzy Rule Interpolation based Q-learning) [18] [21] is a fuzzy rule interpolation (FRI) technique applied Q-learning version, which due to FRI model is suitable for handling continuous state and action spaces. The FRIQ-learning is based on the FIVE (Fuzzy Rule Interpolation based on Vague Environment) interpolation technique [7] [8] [9], which is an application-oriented multidimensional FRI method. The knowledgebase of the system is described by a sparse fuzzy rule-base ($R$), according to following format [18]:

$$r_i: \boldsymbol{If}\ s_1\ is\ S_1^i\ \boldsymbol{And}\ s_2\ is\ S_2^i\ \boldsymbol{And} \dots\ \boldsymbol{And}\ s_n\ is\ S_n^i\ \ \boldsymbol{And}\ a\ is\ A^i\ \boldsymbol{Then}\ \tilde{Q}(\boldsymbol{s}, a) = q^i \tag{1}$$

where $r_i$ $(i \in [1, m])$ is the $i^{th}$ rule in the $m$ sized $R$ rule-base, $\tilde{Q}(\boldsymbol{s}, a)$ is the approximated Q-function by FIVE FRI, $q^i$ is the $i^{th}$ rule consequent, $S_j^i$ $(j \in [1, n])$ is fuzzy set of the $i^{th}$ rule in the $j^{th}$ antecedent dimension, $\boldsymbol{S}$ is the $n$-dimensional observation $(s_1, s_2 \dots s_n \in \boldsymbol{S})$, $s_j$ is the $j^{th}$ dimension of the $\boldsymbol{S}$ state observation, $A^i$ is the fuzzy set of the $i^{th}$ rule in the one-dimensional $U$ action space and $a$ $(a \in U)$ is the executed action.

The a priori knowledge of the system can be defined by human expert, it can also be formalized by fuzzy rule-base ($R_{expert}$) similar to (1) form. The expert determines by this rules which action should be preferred in the corresponding state (as heuristic policy modifier [4]). The form of the $\hat{r}_i$ $(\hat{r}_i \in R_{expert})$ expert rules is the following [15]:

$$\hat{r}_i\text{: }\mathbf{If}\ s_1 \boldsymbol{is}\ \hat{S}_1^i\ \boldsymbol{And}\ s_2 \boldsymbol{is}\ \hat{S}_2^i\ \boldsymbol{And} \dots \boldsymbol{And}\ s_n \boldsymbol{is}\ \hat{S}_n^i\ \boldsymbol{Then}\ \mathrm{a} = \hat{A}^i \tag{2}$$

where $\hat{A}^i$ is the action as rule consequent of the $i^{th}$ $(i \in [1, \hat{m}])$ expert rule, $\hat{S}_n^i = \left[\hat{S}_1^i, \hat{S}_2^i, \dots \hat{S}_n^i\right]$ is the $n$-dimensional state observation, $\hat{m}$ is the number of the expert rule in the $R_{expert}$ rule-base and $\hat{r}_i$ is the $i^{th}$ expert defined rule. The difference compared to (1) is the rule consequent of (2) the is expert determined action (in the specified corresponding state), not is the $q^i$ value as in case of (1) rule form. The expert rule-base can be injected into the system have to determine initial Q-values. The initial Q-values of the expert rule-base will be specified by a Q-value initialization methodology by the following form [15]:

$$\tilde{Q}_{init} = \eta * \frac{g_{max}}{1-\gamma} \text{ if } \gamma < 1 \tag{3}$$

where $\tilde{Q}_{init}$ is the calculated initial Q-value, $g_{max}$ is the possible maximal reinforcement which can be given by the environment, $\gamma$ is the discount factor and $\eta$ is the discount factor of the $\tilde{Q}_{init}$ estimation (see [15] for more details). After the Q-value initialization method the form of the expert rules will be change the following:

$$\hat{r}_i\text{: }\mathbf{If}\ s_1 \boldsymbol{is}\ \hat{S}_1^i\ \boldsymbol{And}\ s_2 \boldsymbol{is}\ \hat{S}_2^i\ \boldsymbol{And} \dots \boldsymbol{And}\ s_n \boldsymbol{is}\ \hat{S}_n^i\ \boldsymbol{And}\ \mathrm{a} = \hat{A}^i\ \boldsymbol{Then}\ \tilde{Q}(\boldsymbol{s}, a) = \tilde{Q}_{init} \tag{4}$$

The expert defined state antecedent and action consequent will be convert to state-action rule antecedent and the estimated $\tilde{Q}_{init}$-value as rule consequent. After the Q-value initialization method the (4) form expert rule-base will be injected into the system. Due to the FIVE FRI the learning phase starts with $2^{n+1}$ (n is the number of the state space dimension) initial rules at the corner of the n + 1-dimensional hypercube. The consequent values of these $r_i^{\square}$ corner rules are 0 $(q^i = 0)$. There may be a case when any expert rules overlap the FRIQ corner rules, this case lead to contradiction, because of same antecedent but different consequent. In this case the system replaces the corresponding corner rules to the expert rules, therefore the $q^i = 0$ will be change to $q^i = \tilde{Q}_{init}$, due to influencing of expert rules.

In the learning iterations the previously merged rule-base will be grow incrementally according to the following updating rule:

$$q_i^{k+1} = \begin{cases} q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) & \text{if } (\boldsymbol{s}, a) = (\boldsymbol{s}^i, a^i) \text{ for some } i, \\ q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) * \left(1/\delta_{v,i}^{\lambda}\right) \Big/ \left(\sum_{i=1}^{m} 1/\delta_{v,i}^{\lambda}\right) & \text{otherwise} \end{cases} \tag{5}$$

where $q_i^k$ is the consequent value of the $i^{th}$ rule in the $k^{th}$ iteration, $(\boldsymbol{s}, a)$ is the given state-action point, $\delta_{v,i}^{\lambda}$ is the scaling distance between the actual observation and the $i^{th}$ rule antecedent and $\Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a)$ can be formalized according to following:

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \alpha * \left(g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a)\right) \tag{6}$$

where $\alpha$ is the learning rate, $\gamma$ is the discount factor, $g(\boldsymbol{s}, a, \boldsymbol{s}')$ is the value of reinforcement for $\boldsymbol{s} \rightarrow \boldsymbol{s}'$ state transition, $\tilde{Q}^k$ is of the $k^{th}$ and $\tilde{Q}^{k+1}$ is of the $(k+1)^{th}$ iteration consequent value approximated by FIVE FRI.

A new rule will be inserted into the initial corner rule-base for the possible rule position if the nearest rule is also far and the Q-update value $(\Delta\tilde{Q})$ is bigger than the Q-update limit $(\varepsilon_Q)$, therefore $\Delta\tilde{Q} > \varepsilon_Q$. The nearest rule means their distance is less then determined distance threshold limit [17]. Otherwise, if the existing a rule near to given rule point and the Q-update value relatively small $\Delta\tilde{Q} < \varepsilon_Q$ , then only the complete rule-base will be updated. This method called incrementally rule-base construction [19]. The learning task will be finished if no more added rule to the incremental rule-base and the $\Delta\tilde{Q}$ values not change significantly. The incrementally built rule-base my contain redundant rules, they can be remove by the rule-base reduction strategies to decrease the of the knowledgebase of the system (see [16] [20] and [22] for more details).

## 3. The suggested structure of the methodology

In the FRIQ-learning [18] [21] and the expert knowledge injected version [15] only the consequent part (Q-value) of the rule-base can be tuning by the update form (5). If a new rule will be inserted to the rule-base, then the consequents of given rules will be calculated by (5) and in turn the antecedent (state-action) part of all rules will be permanent during the whole learning phase. Otherwise, the Q-update value still relatively small then only the consequent part of the rule-base will be updated. In case of inserted new rule by the system, the state-action point (and the related Q-value) probably located in correct position. The possible rule places determined by given observations, the state-action space grid [18] of the FRIQ-learning omitted, therefore the rules can be located anywhere in the state-action space.

In the expert determined a priori rule-base, by the expert defined state antecedent and the action consequent of rules will be change to state-action antecedent and initial Q-value ($\tilde{Q}_{init}$) consequent according to form (4). Regarding the expert rule-base, can be 4 different cases: properly defined rule-base, properly given fragment rule-base, partially incorrect rule-base and completely incorrect rule-base. Properly determined rules mean the given state belong to correct action. Any rules can be specified by the expert, the number of the rules and the type of the rules (correct/incorrect) significantly influence the efficiency of the system (see [15] for more details).

The suggested rule-base optimisation (in other words tuning) methodology is based on the classical gradient descent method. The basic idea is the following: during the rule-base construction the existing rule positions will be change in the in appropriate case with regard to gradient of the Q-function. If there is not exist rule in the examined rule position, the nearest rule is also far compared to given observation then a new rule will be inserted to the actual position of the observation. If the two rules get close to each other during the tuning phase (due to rule position moving), then those will be merged to one cardinal rule.

To apply the gradient descent method, have to determine partial derivatives of the Q-function (5). The Q-function is described by the sparse fuzzy rule-base (due to the applied FIVE FRI model), thus primarily the gradient has to calculate in the rule positions. If already exist rule near to the given observation, then a new rule not will be inserted to the rule position but all of rule positions will be updated according to gradient of the Q-function. Regarding to gradient descent the new rule positions (in the $R$ rule-base) will be change according to the following update form:

$$R_{new} \leftarrow R_{old} - \alpha * \nabla\tilde{Q} \tag{7}$$

Where $R_{new}$ is the new rule positions (for each rule in the complete $R$ rule-base), $R_{old}$ is the old rule positions before the update, $\alpha$ is the step size parameter of the gradient descent and $\nabla\tilde{Q}$ is the gradient (partial derivatives) of the Q-function. Due to the $(n+2)$-variables Q-function, the partial derivatives have to determine for each variable of the function, therefore respect to $\boldsymbol{S}, a$ and $q$, where $\boldsymbol{S}$ is n-dimensional ($s_1, s_2 \dots s_n \in \boldsymbol{S}$). Thus the gradient of the Q-function can be formalized as the following:

$$\nabla\tilde{Q} = \left\{\frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial\boldsymbol{s}}, \frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial a}, \frac{\partial\tilde{Q}(\boldsymbol{s},a)}{\partial q}\right\} \tag{8}$$

The computed gradient of the Q-function in the rule points will be determined the direction of the rules moving in all dimensions $(\boldsymbol{s}, a, q)$. Therefore, in case of no more new incremental rule added to the rule-base then all of rules position will be updated by the gradient descent method. Furthermore, if any rules get close to each other because of the rule position moving, then the closing rules will be merged as the following manner:

$$\begin{gathered} r_{expert} \wedge r_{inserted} \rightarrow r_{expert} \\ r_{expert} \wedge r_{expert} \rightarrow r_{expert} \\ r_{inserted} \wedge r_{inserted} \rightarrow r_{inserted} \end{gathered} \tag{9}$$

The suggestion is the new (merged) rule positions $(\boldsymbol{s}, a)$ determined based on the state-action value average of the closing rules. The closing rules determination is based on rule distance. A rule can be determined as a closing rule of their distance is less than the computed distance thresholds. The distance threshold will be determined for each dimension, if the distance of the given rule is less in each dimension than the computed distance thresholds, the rule can be marked as a closing rule [17].

The diagram of the proposed system is the following:

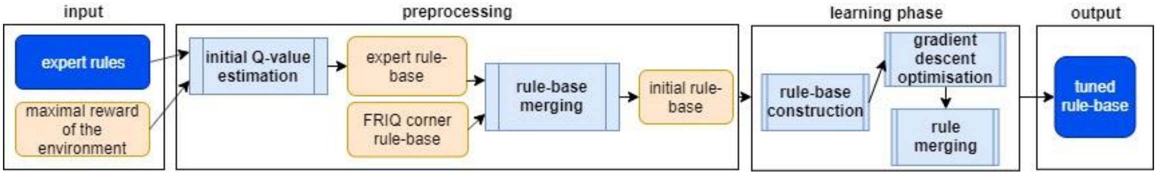

***Figure 1.*** *The proposed structure of the system*

## 4. Conclusions

In the paper a rule-base tuning method is suggested for the expert knowledge extended FRIQ-learning system, which is able for tuning (optimize) the whole expert heuristic included fuzzy rule-base even in the case if it is not correctly defined before the learning phase. In the suggested methodology, the expert can state the a priori knowledge in the form of fuzzy state-action rule-base. Having not entirely correct expert rules in the a priori knowledge, the performance of the learning phase can be degraded. By the proposed tuning method, during the learning phase both the antecedent and consequent parts of the incorrectly given fuzzy rules can be repaired.

## 5. Acknowledgement

The described article was carried out as part of the EFOP-3.6.1-16-2016-00011 “Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialisation” project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

## 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] 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.

[3] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: *Heuristically accelerated reinforcement learning: Theoretical and experimental results*, ECAI. 2012.

[4] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: Accelerating autonomous learning by using heuristic selection of actions, *Journal of Heuristics* 14.2 (2008): pp. 135–168. **https://doi.org/10.1007/s10732-007-9031-5**

[5] Glorennec, P. Y., Jouffe, L.: (1997, July). Fuzzy Q-learning, *In Proceedings of 6th international fuzzy systems conference* (Vol. 2, pp. 659-662). IEEE.

[6] Hindriks, K. V., De Boer, F. S., Van Der Hoek, W., Meyer, J. J. C.: (2000, July). *Agent programming with declarative goals*, In International Workshop on Agent Theories, Architectures, and Languages (pp. 228–243). Springer, Berlin, Heidelberg. **https://doi.org/10.1007/3-540-44631-1_16**

[7] Kovács, Sz., Kóczy, L. T.: 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, pp. 144–149.

[8] Kovács, Sz., Kóczy, L. T.: *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.

[9] Kovács, Sz.: *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.

[10] Kovács, Sz.: Extending the fuzzy rule interpolation“FIVE” by fuzzy observation, Computational Intelligence, *Theory and Applications* (2006): pp. 485–497. **https://doi.org/10.1007/3-540-34783 -6_48**

[11] Matignon, L., Laurent, J. G., Le Fort-Piat, N.: Reward function and initial values: better choices for accelerated goal-directed reinforcement learning, *International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, 2006. **https://doi.org/10.1007/11840817_87**

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

[13] Sutton, R. S., Barto, A. G.: Reinfo*rcement learning: An Introduction*, MIT Press, Cambridge (1998). **https://doi.org/10.1109/TNN.1998.712192**

[14] Tompa, T., Kovács, Sz.: Szakértői heurisztika alkalmazása a FRIQ-learning megerősítéses tanulási módszerben, *Multidiszciplináris Tudományok* 9.4 (2019): pp. 356–368. **https://doi.org/10.35925/j.multi.2019.4.35**

[15] Tompa, T., Kovács, Sz.: Applying expert heuristic as an a priori knowledge for FRIQ-learning, *Acta Polytechnica Hungarica* 17.4 (2020). **https://doi.org/10.12700/APH.17.4.2020.4.2**

[16] Tompa, T., Kovács, Sz.: Clustering-based fuzzy knowledge-base reduction in the FRIQ-learning, Applied Machine Intelligence and Informatics (SAMI), *2017 IEEE 15th International Symposium on. IEEE*, 2017. **https://doi.org/10.1109/SAMI.2017.7880302**

[17] Tompa, T., Kovács, Sz.: Determining the minimally allowed rule-distance for the incremental rule-base contruction phase of the FRIQ-learning, *2018 19th International Carpathian Control Conference (ICCC). IEEE*, 2018. **https://doi.org/10.1109/CarpathianCC.2018.8399677**

[18] Vincze, D., Kovács, Sz.: Fuzzy rule interpolation-based Q-learning, Applied Computational Intelligence and Informatics, 2009. *SACI'09. 5th International Symposium on. IEEE*, 2009. **https://doi.org/10.1109/SACI.2009.5136311**

[19] Vincze, D., Kovács, Sz.: *Incremental rule base creation with fuzzy rule interpolation-based Q-learning*, I. J. Rudas et al. (Eds.), Computational Intelligence in Engineering, Studies in Computational Intelligence, Volume 313/2010, Springer-Verlag, Berlin Heilderberg, 2010, pp. 191-203. **https://doi.org/10.1007/978-3-642-15220-7_16**

[20] Vincze, D., Kovács, Sz.: Rule-base reduction in fuzzy rule interpolation-based Q-learning, *Recent Innovations in Mechatronics (RIiM)* Vol. 2 (2015) No. 1–2. **https://doi.org/10.17667/riim.2015.1-2/10**

[21] Vincze, D.: Fuzzy rule interpolation and reinforcement learning, *15th International Symposium on Applied Machine Intelligence and Informatics (SAMI 2017)*, Herl'any, Slovakia, pp. 173–178. **https://doi.org/10.1109/SAMI.2017.7880298**

[22] Vincze, D., Tóth, A., Niitsuma, M.: Antecedent redundancy exploitation in fuzzy rule interpolation-based reinforcement learning, *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics* (AIM). IEEE, 2020. **https://doi.org/10.1109/AIM43001.2020.9158875**

[23] Watkins, C. J. C. H.: *Learning from delayed rewards*, Ph.D. thesis, Cambridge University, Cambridge, England (1989).