Reinforcement Learning-Based Tuning Algorithm Applied to Fuzzy Identification

Mariela Cerrada[1], Jose Aguilar[1], and André Titli[2]

[1] Universidad de Los Andes,
Control Systems Department-CEMISID, Mérida-Venezuela
{cerradam, aguilar}@ula.ve
[2] DISCO Group, LAAS-CNRS,
Toulouse cedex 4, France
titli@ula.ve

Abstract. In on-line applications, reinforcement learning based algorithms allow to take into account the environment information in order to propose an action policy for the overall optimization objectives. In this work, it is presented a learning algorithm based on reinforcement learning and temporal differences allowing the on-line parameters adjustment for identification tasks. As a consequence, the reinforcement signal is generically defined in order to minimize the temporal difference.

1 Introduction

The Reinforcement Learning (RL) problem has been widely researched an applied in several areas [1, 2, 3, 4, 5, 6, 7, 8]. In dynamical environments, the *learning agent* gets rewards or penalties, according to its performance for learning good actions. In identification problems, information from the environment is needed in order to propose an approximate model, thus, RL can be used for the on-line information taking. Off-line learning algorithms have reported suitable results in system identification, however these results are bounded on the available data, their quality and quantity. In this way, the development of on-line learning algorithms for system identification in an important contribution.

In this work, it is presented an on-line learning algorithm based on RL using the Temporal Difference (TD) method, for identification purposes. Here, the basic propositions of RL with TD are used and, as a consequence, the linear $TD(\lambda)$ algorithm proposed in [1] is modified and adapted for systems identification and the reinforcement signal is generically defined according to the temporal difference and the identification error. Thus, the main contribution of this paper is the proposition of a generic on-line identification algorithm based on RL. The proposed algorithm is applied in the parameters adjustment of a Dynamical Adaptive Fuzzy Model (DAFM) [9], and an illustrative example for time-varying non-linear identification is presented.

J. Wang et al. (Eds.): ISNN 2006, LNCS 3971, pp. 623–630, 2006.

2 Theoretical Background

2.1 Reinforcement Learning and Temporal Differences

RL deals with the problem of learning based on trial and error in order to achieve the overall objective [1]. RL are related to problems where the learning agent do not know what it must do. At time t, $(t = 0, 1, 2, ...)$, the agent receives the *state* S_t and based on this information it choice an *action* a_t. As a consequence, the agent receives a *reinforcement signal or reward* r_{t+1}. In case of the infinite time domain, a *discount* weights the received reward and the *discounted expected gain* is defined as:

$$R_t = r_{t+1} + \mu r_{t+2} + \mu^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \mu^k r_{t+k+1} \tag{1}$$

where μ, $0 \leq \mu \leq 1$, is the *discount rate*, and it determines the current value of the futures rewards.

On the other hand, TD method permits to solve the prediction problem taking into account the difference (error) between two prediction values at successive instants t given by a function P. According to the TD method, the adjustment law for the parameter vector θ of the prediction function $P(\theta)$ in given by the following equation [2]:

$$\theta_{t+1} = \theta_t + \eta(P(x_{t+1}, \theta_t) - P(x_t, \theta_t))\frac{\partial P(x_t, \theta_t)}{\partial \theta} \tag{2}$$

where x_t is a vector of available data at time t and η, $0 \leq \eta \leq 1$, is the learning rate. The term between parenthesis is the *temporal difference* and the equation (2) is the *TD algorithm* and it can be used on-line in a incremental way.

RL problem can be viewed as a prediction problem where the objective is the estimation of the discounted gain defined by equation (1), by using the *TD* algorithm. Let $\hat{R}_t$ be the prediction of R_t, then, from equation (1) and by replacing the real value of R_{t+1} by its estimated value $\hat{R}_{t+1}$, the prediction error between R_t and $\hat{R}_t$ is defined by the equation (3), which describe a temporal difference:

$$\varDelta = R_t - \hat{R}_t = r_{t+1} + \mu \hat{R}_{t+1} - \hat{R}_t \tag{3}$$

By denoting $\hat{R}$ as P and by replacing the temporal difference in (2) by that one defined by the equation (3), the parameters adjustment law is [1]:

$$\theta_{t+1} = \theta_t + \eta(r_{t+1} + \mu P(x_{t+1}, \theta_t) - P(x_t, \theta_t))\frac{\partial P(x_t, \theta_t)}{\partial \theta} \tag{4}$$

2.2 Dynamical Adaptive Fuzzy Models

Without loss of generality, a fuzzy logic model MISO (Multiple Inputs-Single Output), is a linguistic model defined by the following M fuzzy rules:

$$R^{(l)} : IF\ x_1\ is\ F_1^l\ \ AND ...\ AND\ \ x_n\ is\ F_n^l THEN\ \ y\ \ is\ G^l \tag{5}$$

where x_i is a vector of linguistic input on the domain of discourse U_i; y is the linguistic output variable on the domain of discourse V; F_i^l and G^l are fuzzy sets on U_i and V, respectively, $(i = 1, ..., n)$ y $(l = 1, ..., M)$, each one defined by their membership functions.

The DAFM is obtained from the previous rule base (5), by supposing input values defined by fuzzy singleton, gaussian membership functions of the fuzzy sets defined for the fuzzy output variables and the defuzzification method given by center-average method. Then, the inference mechanism provides the following model [9]:

$$y(\underline{X}, t) = \frac{\sum_{l=1}^{M} \gamma^l(u^l, t) \left(\prod_{i=1}^{n} exp\left[-\frac{\left(x_i\ -\ \alpha_i^l(v_i^l,t)\right)^2}{\beta_i^l\ (w_i^l,t)}\right]\right)}{\sum_{l=1}^{M} \left(\prod_{i=1}^{n} exp\left[-\frac{\left(x_i\ -\ \alpha_i^l(v_i^l,t)\right)^2}{\beta_i^l\ (w_i^l,t)}\right]\right)} \tag{6}$$

where $\underline{X} = (x_1\ x_2\ ...\ x_n)^T$ is a vector of linguistic input variables x_i at time t; $\alpha(v, t_j)$, $\beta(w, t_j)$ and $\gamma(u, t_j)$ are time-depending functions; v_i^l y w_i^l are parameters associated to the variable x_i in the rule l; u^l is a parameter associated to the center of the output fuzzy set in the rule l.

Definition 1. Let $x_i(t_j)$ be the value of the input variable x_i to the DAFM at time t_j to obtain the output $y(t_j)$. The generic structure of the functions $\alpha_i^l(v_i^l, t_j)$, $\beta_i^l(w_i^l, t_j)$ and $\gamma^l(u^l, t_j)$ in equation (6), are defined by the following equations [9]:

$$\alpha_i^l(v_i^l, \overline{x}_i(t_j)) = v_i^l \frac{\sum_{k=j-\delta_1}^{j} (x_i(t_k))}{\delta_1 + 1}; \quad \delta_1 \in \aleph \tag{7}$$

$$\beta_i^l(w_i^l, \sigma_i^2(t_j)) = w_i^l * (\frac{\sum_{k=j-\delta_1}^{j} (x_i(t_k) - \overline{x}_i(t_k))^2}{\delta_1 + 1} + \epsilon); \quad \epsilon \in \Re \tag{8}$$

$$\gamma^l(u^l, \overline{y}(t_j)) = u^l \frac{\sum_{k=j-\delta_2}^{j-1}\ y(t_k)}{\delta_2}; \quad \delta_2 \in \aleph \tag{9}$$

3 RL-Based Identification Algorithm for DAFM

In this work, the fuzzy identification problem is solved by using the weighted identification error as a prediction function in the RL problem, and by suitably defining the reinforcement value according to the identification error. Thus, the minimization of the prediction error (3) drives to the minimization of the identification error. The *critic* (learning agent) is used in order to predict the performance on the identification as an approximator of the system's behavior. The prediction function is defined as a function of the *identification error* $e(t, \theta_t) = y(t) - y_e(t, \theta_t)$, where $y(t)$ denotes the real value of the system output at time t and $y_e(t, \theta_t)$ denotes the estimated value given by the identification model by using the available values of θ at time t.

Let P_t be the proposed non-linear prediction function in equation (10):

$$P(x_t, \theta_t) = \frac{1}{2} \sum_{k=t-K}^{t} (\mu\lambda)^{t-k} e^2(k, \theta_t) \tag{10}$$

where $e(t, \theta_t) = y(t) - y_e(t, \theta_t)$ defines the identification error and K defines the size of the time interval. Then:

$$\frac{\partial P(x_t, \theta_t)}{\partial \theta} = \sum_{k=t-K}^{t} (\mu\lambda)^{t-k} e(k, \theta_t) \frac{\partial e(k, \theta_t)}{\partial \theta} \tag{11}$$

By replacing (11) into (4), the following learning algorithm for the parameters adjustment is obtained:

$$\theta_{t+1} = \theta_t + \eta(r_{t+1} + \mu P(x_{t+1}, \theta_t) - P(x_t, \theta_t)) \sum_{k=t-K}^{t} (\mu\lambda)^{t-k} e(k, \theta_t) \frac{\partial e(k, \theta_t)}{\partial \theta} \tag{12}$$

The function $P(x_{t+1}, \theta_t)$ in equation (13) is obtained from (10) and, finally, by replacing (13) into (12), the proposed learning algorithm is given.

$$P(x_{t+1}, \theta_t) = \frac{1}{2} e^2(t+1, \theta_t) + \mu\lambda P(x_t, \theta_t) \tag{13}$$

In the prediction problem of the discounted expected gain R_t, a good estimation of R_t given by $\hat{R}_t$ is expected; that implies $P(x_t, \theta_t)$ goes to $r_{t+1} + \mu P(x_{t+1}, \theta_t)$. This condition is obtained from equation (3). Given that the prediction function is the weighted sum of the square identification error $e^2(t)$, then it is expected that:

$$0 \leq r_{t+1} + \mu P(x_{t+1}, \theta_t) < P(x_t, \theta_t) \tag{14}$$

On the other hand, a suitable adjustment of identification model means that the following condition is accomplished:

$$0 < P(x_{t+1}, \theta_t) < P(x_t, \theta_t) \tag{15}$$

The reinforcement r_{t+1} is defined in order to accomplish the expected condition (14) and taking into account the condition (15). Then, by using equations (10) and (13):

$$\begin{aligned} r_{t+1} &= 0 \quad if \quad P(x_{t+1}, \theta_t) \leq P(x_t, \theta_t) \\ r_{t+1} &= -\frac{1}{2} \mu e^2(t+1, \theta_t) \quad if \quad P(x_{t+1}, \theta_t) > P(x_t, \theta_t) \end{aligned} \tag{16}$$

In this way, the identification error into the prediction function $P(x_{t+1}, \theta_t)$, according to the equation (13), is rejected by using the reinforcement in equation (16). The learning rate η in (12) is defined by the equation (17). Parameters μ

and λ can depend on the system dynamic: small values in case of slow dynamical systems, and values around 1 in case of fast dynamical systems.

$$\eta(t) = \frac{\eta(t-1)}{\rho + \eta(t-1)}, 0 < \rho < 1 \tag{17}$$

The proposed identification learning algorithm can be studied like a descent-gradient method with respect to the parametric predictive function P. In the descent-gradient method, the objective is to find the minimal value of the error measure on the parameters space, denoted by $J(\theta)$, by using the following algorithm for the parameters adjustment:

$$\theta_{t+1} = \theta_t + \Delta\theta_t = \theta_t + 2\alpha(E\{z|x_t\} - P(x_t, \theta))\nabla_\theta P(x_t, \theta) \tag{18}$$

In this case, an error measure is defined as:

$$J(\theta, x) = (E\{z|x\} - P(x, \theta))^2 \tag{19}$$

where $E\{z|x\}$ is the expected value of the real value z, from the knowledge of the available data x. In this work, the learning algorithm (12) is like a learning algorithm (18), based on the descent-gradient method, where $r_{t+1} + \mu P(x_{t+1}, \theta_t)$ is the expected value $E\{z|x\}$ in (19). By appropriate selecting r_{t+1} according to (16), the expected value in the learning problem is defined in two ways:

$$E\{z|x\} = \mu P(x_{t+1}, \theta_t) \; if P(x_{t+1}, \theta_t) \leq P(x_t, \theta_t) \tag{20}$$

or

$$E\{z|x\} = \mu^2 \lambda P(x_t, \theta_t) \; if P(x_{t+1}, \theta_t) > P(x_t, \theta_t) \tag{21}$$

Then, the parameters adjustment is made on each iteration in order to attain the expected value of the prediction function P according to the predicted value of $P(x_{t+1}, \theta_t)$ and the real value $P(x_t, \theta_t)$. In both of cases, the expected value is minor than the obtained real value $P(x_t, \theta_t)$ and the selected value of r_{t+1} defines the magnitude of the defined error measure.

4 Illustrative Example

This section shows an illustrative example applied to fuzzy identification of time-varying non-linear systems by using the proposed on-line RL-based identification algorithm in order to adjust the parameters v_i^l, w_i^l and u^l of the DAFM described in section 2.2. The performance of the fuzzy identification is evaluated according to the identification relative error ($e_r = \frac{y(t) - y_e(t)}{y(t)}$) normalized on $[0, 1]$.

The system is described by the following difference equation:

$$y(t+1) = \frac{y(t)y(t-1)y(t-2)u(t-1)(y(t-2)-1) + u(t)}{a(t) + y(t-2)^2 + y(t-1)^2} = g[.] \tag{22}$$

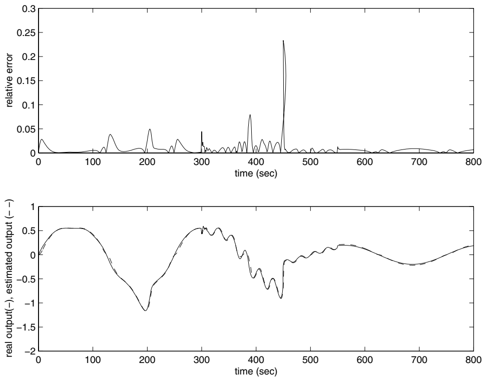

Fig. 1. Fuzzy identification using off-line tuning algorithm

where $a(t) = 1 + 0.1\sin(2\pi t/100)$. In this case, the unknown function $g = [.]$ is estimated by using the DAFM and, additionally, a sudden change on $a(t)$ is proposed by setting $a(t) = 5$, $t > 450$.

Figure 1 shows the performance of the DAFM using the off-line gradient-based tuning algorithm with initial conditions on the interval $[0, 1]$ and using the input signal (23). After an extensive training phase, the fuzzy model with $M = 8$ is chosen.

$$u(t) = \begin{cases} 1.5 + (0.8\sin(2\pi t/250) + 0.2\sin(2\pi t/25)) & if \quad 301 < t < 550 \\ \sin(2\pi t/250) & if \quad otherwise \end{cases} \tag{23}$$

In the following, fuzzy identification performance by using the proposed RL-based tuning algorithm is presented. Here, $\lambda = \mu = 0.9$, $K = 5$ and the learning rate is set up by the equation (17) with $\eta(0) = 0.01$. After experimental proofs, the performance approaching the accuracy obtained from off-line adjustment is obtained with $M = 20$, figure 2 shows the tuning algorithm performance. However, a good performance is also obtained with $M = 8$. Table 1 shows the comparative values related to the RMSE. Figure 3, shows the algorithm sensibility according to the initial conditions and figure 4 shows the algorithm performance under changes on the internal dynamics by taking $a(t) = 1 + 0.3\sin(2\pi t/10)$.

Table 1. Comparison between the on-line proposed algorithm and off-line tuning

M	RMSE on-line	RMSE off-line
8	0.0323	0.0156
10	0.0339	0.0837
15	0.0339	0.0308
20	0.0205	0.1209

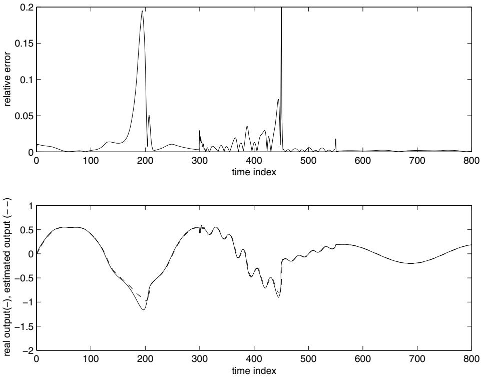

Fig. 2. Fuzzy identification using RL-based tuning algorithm. Initial conditions on $[0.5, 1.5]$.

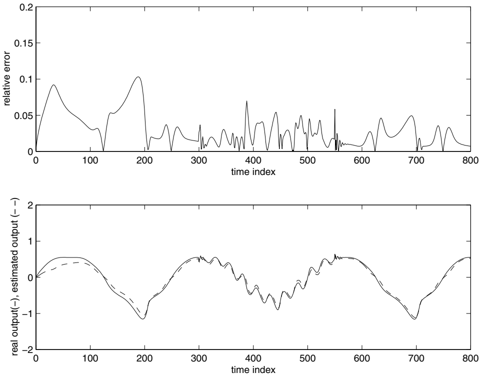

Fig. 3. Fuzzy identification using RL-based tuning algorithm. Initial conditions on $[0, 1]$.

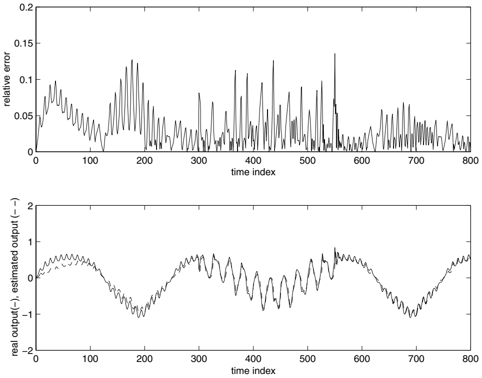

Fig. 4. Fuzzy identification using RL-based tuning algorithm

The previous tests show the performance and the sensibility of the proposed on-line algorithm is adequate in terms of the initial conditions of the DAFM parameters, changes on the internal dynamic and changes on the inputs signal. Table 1 also shows the number of rules M do not strongly determines the global performance of the proposed on-line algorithm.

5 Conclusions

In this work, an on-line tuning algorithm based on reinforcement learning for identification problem has been proposed. Both the prediction function and the reinforcement signal have been defined by taking into account the identification error and the obtained algorithm can be studied like a descend-gradient-based method. In order to show the algorithm performance, an illustrative example related to time-varying non-linear system identification using a DAFM has been developed. The performance of the on-line algorithm is adequate in terms of the main aspects to be taken into account in on-line identification: the initial conditions of the model parameters, the changes on the internal dynamic and the changes on the input signal. This one highlights the use of the on-line learning algorithms and the proposed RL-based on-line tuning algorithm could be an important contribution for the system identification in dynamical environments with perturbations, for example, in process control area.

References

1. Sutton, R., Barto, A.: Reinforcement Learning. An Introduction. The MIT Press, Cambridge (1998)
2. Sutton, R.: Learning to Predict by the Methods of Temporal Differences. Machine Learning **3** (1988) 9-44
3. Miller, S., Williams R.: Temporal Difference Learning: A Chemical Process Control Application. In Murray A., ed.: Applications of Artificial Neural Networks. Kluwer, Norwell (1995)
4. Singh, S., Sutton, R.: Reinforcement Learning with Replacing Eligibility Traces. Machine Learning **22** (1995) 123-158
5. Schapire, R., Warmuth, M.: On the Worst-case Analysis of Temporal Difference Learning Algorithmes. Machine Learning **22** (1996) 95-121
6. Tesauro, G.: Temporal Difference Learning and TD-Gammon. Communications of the Association for Computing Machinery **38(3)** (1995) 58-68
7. Si, J., Wang Y.: On Line Learning Control by Association and Reinforcement. IEEE Transactions on Neural Networks **12(2)** (2001) 264-276
8. Van-Buijtenen, W., Schram, G., Babuska, R., Verbruggen, H.: Adaptive Fuzzy Control of Satellite Attitude by Reinforcement Learning. IEEE Transactions on Fuzzy Systems **6(2)** (1998) 185-194
9. Cerrada, M., Aguilar, J., Colina, E., Titli A.: Dynamical Membership Functions: An Approach for Adaptive Fuzzy Modeling. Fuzzy Sets and Systems **152** (2005) 513-533