Adaptive Fuzzy Control of Satellite Attitude by Reinforcement Learning

Walter M. van Buijtenen, Gerard Schram, Robert Babuška, and Henk B. Verbruggen

***Abstract*— The attitude control of a satellite is often characterized by a limit cycle, caused by measurement inaccuracies and noise in the sensor output. In order to reduce the limit cycle, a nonlinear fuzzy controller was applied. The controller was tuned by means of reinforcement learning without using any model of the sensors or the satellite. The reinforcement signal is computed as a fuzzy performance measure using a noncompensatory aggregation of two control subgoals. Convergence of the reinforcement learning scheme is improved by computing the temporal difference error over several time steps and adapting the critic and the controller at a lower sampling rate. The results show that an adaptive fuzzy controller can better cope with the sensor noise and nonlinearities than a standard linear controller.**

***Index Terms*— Autonomous control, neuro-fuzzy control, reinforcement learning, satellite attitude control.**

I. Introduction

RECENT advances in the technology have led to higher requirements on the performance of control systems. Since many problems are inherently nonlinear and exhibit uncertainty that cannot be modeled in the stochastic framework, new methods are being sought to cope with these phenomena. Methods based on fuzzy sets and fuzzy logic proved to be suitable for designing nonlinear controllers and for dealing with nonprobabilistic uncertainty [1]. There is a general agreement that fuzzy logic provides a suitable framework for the incorporation of *a priori* knowledge in the control design. One of the problems in fuzzy control remains the tuning and adaptation of the controller. Recently, much research has been focused on the combination of fuzzy control with learning algorithms originating from the field of neural networks. These methods are usually referred to as neuro-fuzzy control (see, for instance, [2] among many other references). One of the neuro-fuzzy control techniques is based on a combination of fuzzy modeling and control structures with reinforcement learning (RL) [3], [4]. The main advantage of reinforcement learning is that no model of the process is required for the adaptation of the controller [5]. The main idea of the RL technique is learning through exploration of the space of possible control actions. Actions that result in a good performance are "rewarded" and actions leading to a poor performance are "penalized."

This paper presents an application of reinforcement learning to the adaptation of a nonlinear fuzzy controller for satellite attitude control. The satellite is designed for astronomical observations of infrared light-emitting celestial objects and uses a star tracker as the primary attitude sensor. Usually, the satellite is equipped with a linear proportional derivative (PD) controller in combination with a Kalman filter for the estimation of the noise-free sensor attitude error and error rate (LQG control). Since the star-tracker signal is corrupted by a significant level of noise of a nonstochastic nature, the attitude control of the satellite is characterized by a limit cycle that cannot be eliminated by using the linear Gaussian control. Moreover, the design of the LQG controller is time consuming since it involves modeling of the satellite and of the sensors. Moreover, the obtained solution is specific for the given noise characteristics. As the noise characteristics vary for different objects (weak and strong stars), adaptation features are desirable in order to maintain satisfactory performance under all conditions.

Berenji *et al.* [4] demonstrated that adaptive fuzzy control by means of RL can be applied to attitude control. Our approach is also based on reinforcement learning, but there are several important differences between the scheme of Berenji *et al.* called GARIC and the approach presented in this paper.

- While GARIC uses a binary reinforcement signal indicating either successful performance or a failure, our approach is based on a real-valued reinforcement signal that indicates a degree of satisfaction of the control goals. The control goals are described by using fuzzy sets for the attitude error and error rate. This approach is more flexible, since it allows for a more accurate formulation of the control criteria and aggregation of different goals. The continuous reinforcement signal also provides more detailed information about the controller performance. However, as indicated in Section IV, the usual RL technique does not perform well for continuous reinforcement signals. Therefore, a modified RL scheme is proposed in this paper.
- The fuzzy controller structure differs from the one employed in GARIC. The Takagi–Sugeno (TS) rules with constant consequent functions are used in this paper to implement both the critic and the controller while the controller in GARIC is based on linguistic rules. The TS scheme is computationally more efficient, equally transparent to interpretation of both the initial and the adapted controller and provides intuitively more understandable and consistent results than those presented in [4].
- Since there is no failure situation defined, no backup controller is used. In our setting, the adaptive fuzzy con-

Manuscript received February 26, 1996; revised February 10, 1997.

The authors are with the Department of Electrical Engineering, Delft University of Technology, Delft, 2600 GA, The Netherlands.

Publisher Item Identifier S 1063-6706(98)00808-X.

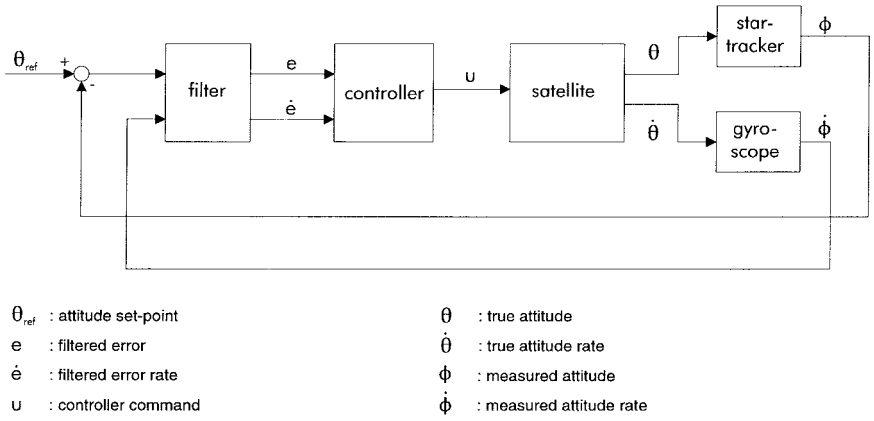


Fig. 1. The control diagram of the satellite.

troller is initialized using the parameters of a reasonably performing linear controller (the constant consequents of the TS rules can be exactly computed to achieve the desired linear control surface). This controller maintains stable control without the need of switching to a backup controller, and further improves its performance by means of adaptation.

- The update law for the controller parameters is based on the original approach of Barto [5] instead of using an approximation of the process Jacobian as in [4].

The rest of this paper is organized as follows. Section II describes the satellite attitude control problem. Section III gives the necessary background of the applied reinforcement learning technique. Section IV describes the adaptive control scheme of the satellite and presents the results. Section V concludes the paper.

II. Satellite Attitude Control

During observation of a celestial body with a telescope, the attitude of the satellite is controlled by means of a linear PD controller. A star tracker is used as an attitude sensor. The attitude rate is measured by a gyroscope and a Kalman filter is employed to estimate the noise-free attitude error and error rate (see Fig. 1).

The star tracker consists of a coupled-charge device (CCD) camera. Subpixel resolution is obtained by defocussing the stellar image on an array of 3×3 pixels. When the star moves in the field of view as a consequence of the satellite motion and attitude control, the star tracker automatically adapts the selection of the pixel array in order to keep the pattern of 3×3 pixels centered on the star image. In Fig. 2, the selection of a different set of pixels is illustrated (from pixel set a to b). Each pixel can be considered as a separate detector, with its own characteristics (sensitivity to light). Differences between the characteristics of the pixels in the array cause errors in the measured position of the star, and hence in the measured attitude. For a particular position and particular properties of the star light, the error is constant and it is called the "bias." When the center of the star image crosses the border between two adjacent pixels, a different set of pixels is selected, causing a discontinuity in the bias. These discontinuities along with the measurement noise deteriorate the control performance and result in a limit cycle.

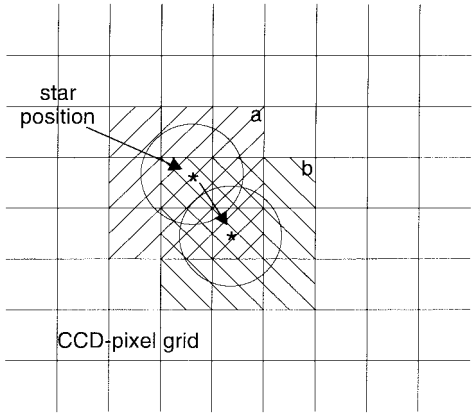
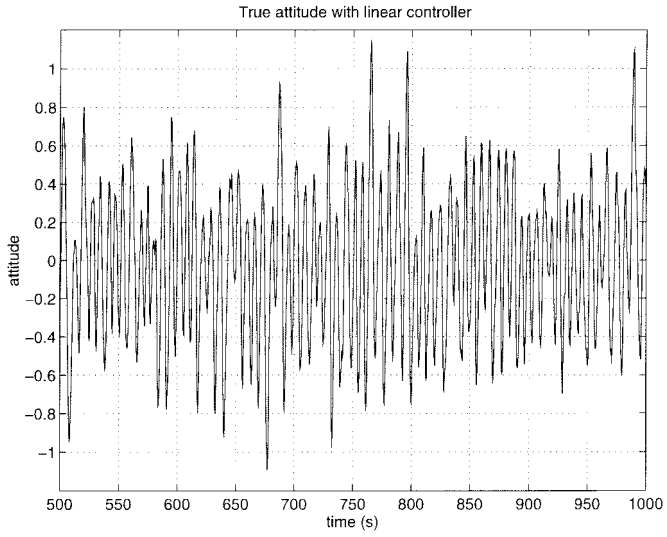


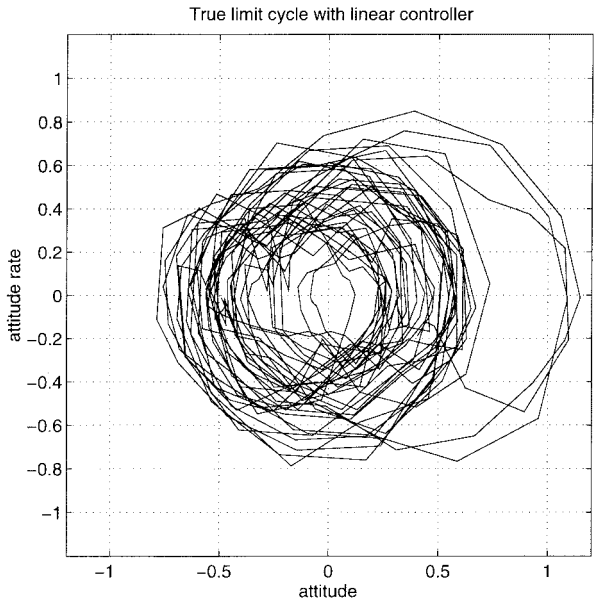
Fig. 2. Tracking of a star by selecting a different set of pixels.

The overall system, containing a conventional PD controller with factory settings was simulated over a time interval of 1000 s. In this simulation, the satellite is controlled at an attitude corresponding to the border of two adjacent pixels. This is called the $+2/-2$ bias situation, where the output of the star tracker is two arcseconds too high for angles larger than the set point and two arcseconds too low for other angles. The set point used in the simulations is zero arcseconds. In Fig. 3(a), the corresponding attitude of the satellite is plotted in time domain. Fig. 3(b) shows the limit cycle in the phase plane. The calculated root-mean-square (rms) attitude deviation is 0.41 arcseconds.

In a previous study, Schram *et al.* [6] proposed to tune a conventional PD controller of the satellite by reinforcement learning. However, the linear nature of the controller does not allow for a significant improvement of the performance. It is expected that better controller performance can be achieved when control actions are locally adjusted, resulting in a nonlinear control law. In this paper, the nonlinear controller is implemented as a TS fuzzy controller. This structure allows for the initialization of the fuzzy controller by a reasonably functioning linear controller and for further improvement of the controller performance by means of adaptation. The adapted rule base remains transparent and interpretable. A



(a)



(b)

Fig. 3. Attitude limit cycle in the $+2/-2$ bias situation.

fuzzy performance measure is used as the reinforcement signal based on a noncompensatory aggregation of two control subgoals for the attitude error and for its rate.

III. Reinforcement Learning

The term reinforcement learning (RL) refers to a family of algorithms inspired by human and animal learning. The objective of RL is to discover a control policy, i.e., a mapping from states to control actions. In RL, there is no direct evaluation of the selected control action. Instead, an indirect evaluation is received in terms of (dis)satisfaction of the control objectives. The goal of RL is to discover such control policy that maximizes the reinforcement received.

The reinforcement signal is often defined as a scalar value, which is usually -1 to express a failure and 1 or 0 to indicate a success. Also a more detailed (continuous) degree

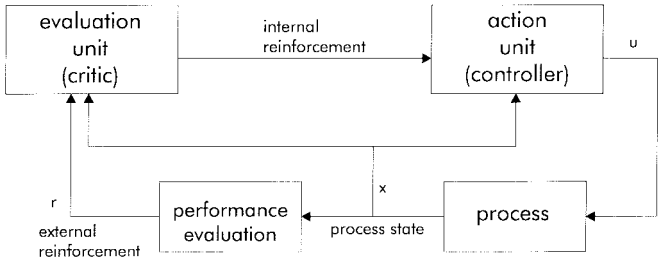


Fig. 4. A general reinforcement learning scheme.

of failure or success can be used, as is the case in our paper. Since there is no teacher or supervisor who could evaluate the selected control actions, RL techniques use an internal evaluator called the critic. The RL method searches for the best actions by exploration (deliberate modification of the control actions computed by the controller) and by evaluating the consequences of these modifications.

A general scheme realizing this type of learning is given in Fig. 4. This scheme, also called policy iteration, was proposed by Barto [5] and generalized by Anderson [7]. It consists of two units: the critic (evaluation unit) and the controller (action unit). The critic predicts the expected future reinforcement the process will receive as being in the state x and following the current controller policy. The action unit consists of the controller and a stochastic exploration module (not shown in the figure). The stochastic exploration is needed to explore the space of possible actions.

A. The Critic

The task of the critic is to predict the future system performance. This prediction is needed to obtain a more informative signal (internal reinforcement), which can be used to adapt the critic and the controller.

In simple RL problems where the reinforcement signal refers only to the last applied control action, it is sufficient that the critic predicts one step ahead (immediate RL). In more complex dynamic learning tasks, the control actions cannot be judged individually because of the dynamics of the process. The reinforcement signal then refers to an action that has been taken in the past. It is not known which particular action is responsible for the given state, which leads to the credit assignment problem also called delayed RL [5].

In order to solve the credit assignment problem, the critic is trained to predict the expected sum of future external reinforcement signals. Hence, in the delayed RL, the critic acts as a multistep predictor, whereas in the case of immediate RL, it is a single-step predictor. The sum of future reinforcements which the critic learns to predict is given by

$$V(k) = \sum_{i=k}^{\infty} \gamma^{i-k} r(i+1) \tag{1}$$

where $\gamma \in [0, 1\rangle$ is an exponential discounting factor, r is the external reinforcement signal, k denotes a discrete time instant, and $V(k)$ is the discounted sum of future reinforcements (also called the value function). The critic is usually implemented as a nonlinear function approximator such as a neural network

[7] or a fuzzy system [4]. In order to derive an update law for the critic parameters, let us denote $\hat{V}(k)$ the prediction of $V(k)$ computed by the critic. By rewriting (1) as

$$V(k) = r(k+1) + \gamma V(k+1) \tag{2}$$

we can derive a prediction error Δ resulting from an incorrect prediction $\hat{V}(k)$ as

$$\Delta = [r(k+1) + \gamma \hat{V}(k+1)] - \hat{V}(k). \tag{3}$$

Since the prediction error Δ is computed from two consecutive values $\hat{V}(k)$ and $\hat{V}(k+1)$, it is called the *temporal difference* (TD) [8]. Note that both $\hat{V}(k)$ and $\hat{V}(k+1)$ are known at time k since $\hat{V}(k+1)$ is a prediction obtained for the current process state. In the literature, the temporal difference error is also referred to as the internal reinforcement signal [5]. The term between the square brackets represents a training target of the value function. It contains the immediate system payoff (reinforcement), which acts as a reference. Hence, the temporal difference can directly be used to adapt the critic by

$$w(l+1) = w(l) + \alpha_w \Delta \frac{\partial \hat{V}(k)}{\partial w} \tag{4}$$

where w are adaptable parameters of the critic, Δ is the temporal difference, and $\alpha_w > 0$ is the learning rate of the critic. $\partial \hat{V}(k)/\partial w$ is a partial derivative of the critic output with respect to its parameters. The argument l denotes the lth iteration of the parameter update. In general, l may be different from k when the critic is adapted at a lower rate than the sampling rate (see Section IV). Note that for $\gamma \to 0$ a single-step predictor of the immediate RL is obtained as a special case of (3).

B. The Controller

When the critic has learned to predict the future system's performance, the controller can be trained in order to establish an optimal mapping between the system states and the control actions in the sense that the criterion (1) is maximized. Three common approaches can be found in the literature.

The first approach is called Q-learning [9]. This method selects a control action (out of a finite set of actions) that most increases the performance criterion. This type of learning approximates dynamic programming. The second approach uses gradient information, see for instance [3]. The gradient of the estimated criterion with respect to the controller command is calculated, assuming a differentiable criterion function. The parameters of the controller are adapted in the direction of the positive gradient. The third approach uses the temporal difference between the expected performance and the measured performance [5]. When a stochastically modified control action results in a positive difference, i.e., the resultant performance is better than was predicted, the controller has to be "rewarded," and vice versa. In this paper, the last approach is used since it does not require the derivative of the process. Q-learning is not suitable in our case since continuous control actions are used.

Assume that the critic is already able to accurately predict the future reinforcements. Given a certain state, a controller action u is calculated using the current controller. This action is not applied to the process, but it is stochastically modified in order to perform exploration. The actual action u', which is applied to the process, is constructed by adding a random value from $\mathrm{N}(0, \sigma)$ to u. After the action u' is sent to the process, the internal reinforcement signal is calculated. For the delayed RL scheme, the following controller update law can be derived [2]:

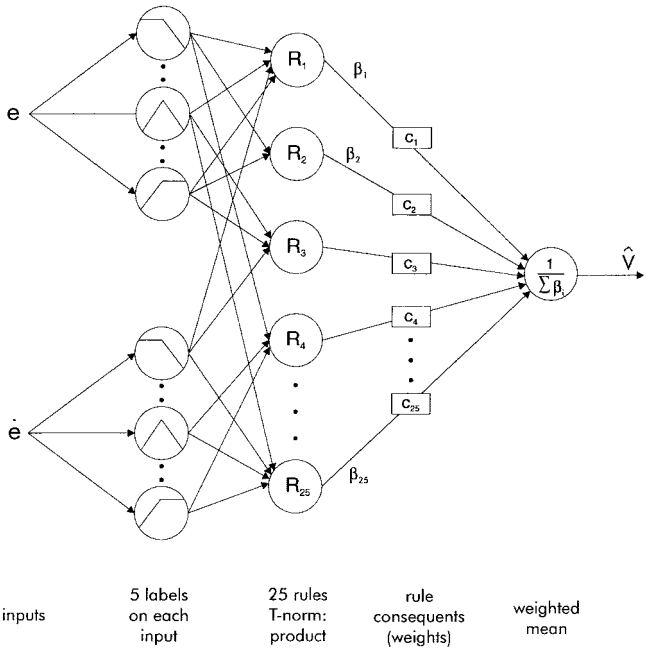


Fig. 5. A neuro-fuzzy implementation of the critic.

$$v(l+1) = v(l) + \alpha_v \left[\frac{u'(k) - u(k)}{\sigma} \right] \Delta \frac{\partial u(k)}{\partial v} \tag{5}$$

where v are the adaptable parameters of the controller, $\alpha_v > 0$ is the learning rate of the controller, and σ is the variance of the normal distribution of the stochastic action modifier. The term between the square brackets is a normalized difference between the actual and the computed control action. The adaptation of the controller relies on the accurate prediction of the critic, therefore, it is necessary to train the critic first or to let the critic adapt with a higher learning rate than the controller. In most applications, first the critic is trained with a fixed controller and without exploration. After the critic has learned to predict the current controller's performance, exploration is enabled and the adaptation of the controller starts with a learning rate smaller than that of the critic.

IV. Adaptive Fuzzy Control of the Satellite

Simulation experiments with the satellite showed that the result of a controller action can already be detected at a next time step as long as the exploration signal has a large magnitude. This is also the reason why experiments with immediate RL were done [10]. However, the learning algorithm with immediate reinforcement diverges due to the Kalman filter (from the LQG scheme), which is designed for specific noise levels and remains implemented in the control scheme. When the controller is changed by the learning algorithm,

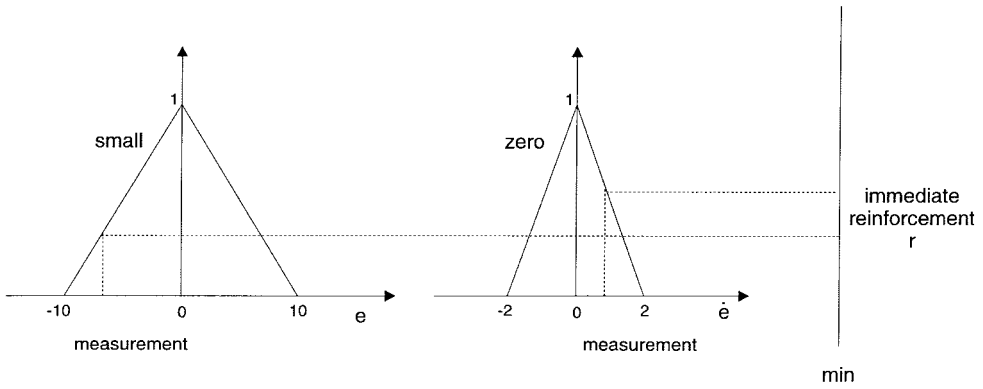


Fig. 6. Fuzzy sets for computing the immediate reinforcement.

the designed filter is no longer suitable and the estimation of the noise-free error and error rate deteriorates. Hence, for further reduction of the limit cycle and convergence of the learning algorithm, a filter is needed which is independent of the controller performance. Experiments showed that a low-pass fourth-order Butterworth filter (of the same order as the Kalman filter) suffices for both the error and the error rate. The cutoff frequency of the filter was chosen according to the limit cycle frequency and the noise properties. This type of filter introduces a time delay and, as a consequence, the immediate RL is no longer appropriate. Thus, delayed RL has been applied as described below.

A. The Critic Structure

The critic is implemented by TS fuzzy rules with constant consequents [11]. It has two inputs: the estimated (filtered) attitude error and error rate and a single output: the predicted discounted sum of future reinforcements. Five linguistic terms (labels) are defined for each antecedent variable. Fig. 8(a) shows the membership functions, which are chosen to be triangular, symmetrical, and uniformly spread over the input domain. The sum of the membership functions equals one for each domain element. The domains for the membership functions are determined according to the amplitude of the limit cycle and the maximum error rate encountered in simulations with the LQG controller. The critic is trained by adapting only the consequents of the rules. The membership functions in the antecedents of the rules are fixed. The output of the controller is a weighted mean of the individual rule consequents

$$\hat{V} = \frac{\sum_{i=1}^{N_r} \beta_i c_i}{\sum_{i=1}^{N_r} \beta_i} \tag{6}$$

where N_r is the number of rules, c_i is the ith consequent, and β_i is the degree of fulfillment of the ith rule. The degree of fulfillment of each rule is computed using the product operator. With this setting, the TS structure can be seen as a look-up table with linear interpolation. In Fig. 5, a possible neuro-fuzzy interpretation of the critic structure is given (similar to [4]). Note, however, that in order to apply the parameter update formulas (4) and (5), this neural interpretation is not necessary.

The update law for the rule consequents of the critic derived from (4) and (6) is given by

$$c_i(l+1) = c_i(l) + \alpha_w \Delta \beta_i(k). \tag{7}$$

Note that when the product operator is used and the fuzzy sets form a partition, the sum of the degrees of fulfillment over all rules equals one ($\sum_{i=1}^{N_r} \beta_i = 1$) and the partial derivative $\partial V(k)/\partial c_i$ reduces to $\beta_i(k)$.

A fuzzy measure of performance is used to compute the external reinforcement signal r. The control goal is to reduce the limit cycle, i.e., to keep the attitude error low (not necessarily zero) and simultaneously keep the error rate as close as possible to zero. Note that because of the sensor noise, it cannot be expected that the attitude error will be exactly zero. In fact, a small error can be tolerated as long as the satellite stays still or remains *slowly* moving in a limit cycle with a small amplitude. It is well known that this kind of performance specification can be realized in a flexible manner by means of fuzzy goals (and constraints) [12], [13]. In our case, the goal can be formulated as "keep the attitude error small and the error rate close to zero." The meaning of the linguistic terms "small" and "zero" is defined by means of suitably chosen membership functions in the domains of the error and error rate (see Fig. 6). Since no compensation between the two goals is allowed, the logical conjunction operator ("and") is implemented as the minimum operator. In terms of membership degrees, the reinforcement signal r is computed by

$$r = \min[\mu_{\text{Small}}(e), \mu_{\text{Zero}}(\dot{e})]. \tag{8}$$

The shape and the width of the membership functions can be modified to tune the controller's performance.

B. The Fuzzy Controller Structure

The fuzzy controller is also implemented by TS rules with constant consequents and it uses the same inputs and membership functions as the critic. The output of the controller represents the torque command u and is computed as a weighted mean of the individual rule consequents (6). As in the critic, the antecedent fuzzy sets are fixed, only the consequent

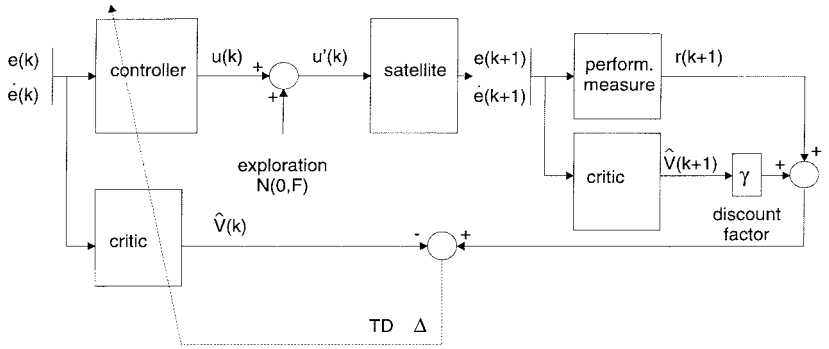


Fig. 7. Delayed reinforcement scheme for the satellite attitude control: e is the error, $\dot{e}$ the error rate, u is the control action, N is the Gaussian exploration noise, u' is the modified control action, r the immediate reinforcement, Δ the temporal difference, and $\hat{V}$ is the criterion prediction.

parameters are updated by

$$b_i(l+1) = b_i(l) + \alpha_v \text{sign}\{[u'(k) - u(k)]\Delta\}\beta_i(k) \qquad (9)$$

where b_i is the ith rule consequent of the controller. Equation (9) is derived from (5) and (6), taking only the sign of the exploration and of the temporal difference in order to prevent relatively large adaptation steps in the wrong direction caused by a bad prediction of immediate reinforcement. This results in a more stable way of adapting the consequents. In Fig. 7 the delayed RL scheme is illustrated. Note that the two critic blocks represent the same critic.

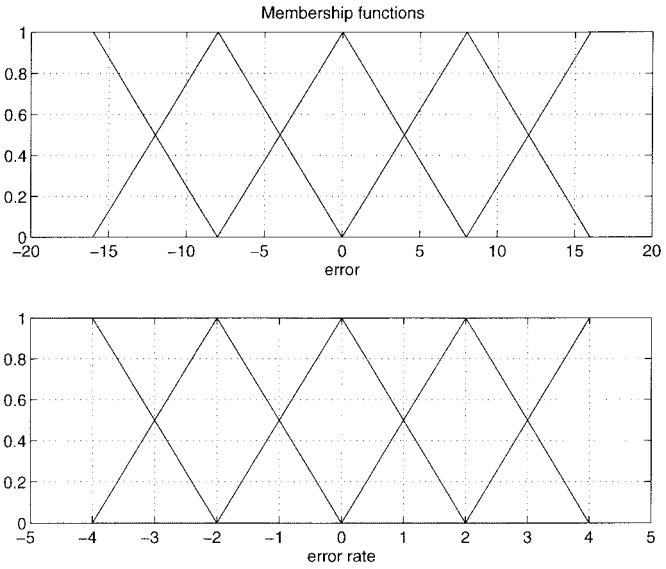
The fuzzy controller is initialized as a linear PD controller with an acceptable performance. In Fig. 8(a) the membership functions and in Fig. 8(b) the control surface is plotted. In Table I the initial rule base of the fuzzy controller is given.

The initial performance of the fuzzy controller has an rms attitude error of 3.43 arcseconds and an rms error of the attitude rate 2.04 arcseconds/s. The satellite is simulated in the $+2/-2$ bias situation.

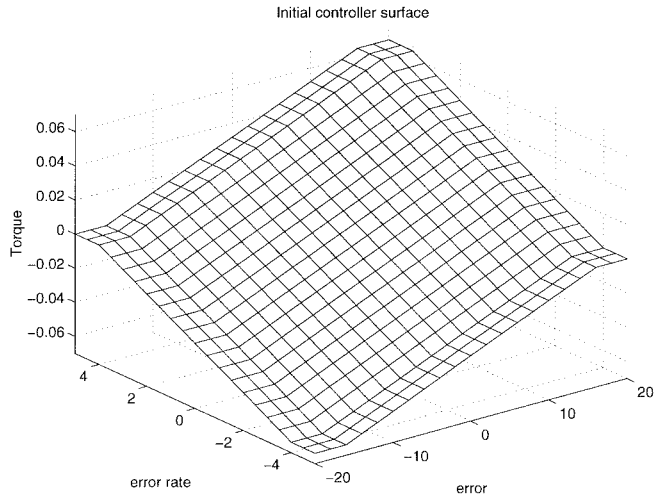
In the learning scheme, both the critic and the controller are adapted simultaneously and the controller is adjusted with information given by the critic. Therefore, the learning rate α_v of the controller update law (9) has a lower value than the learning rate α_w in the update law (7) of the critic: $\alpha_v = 0.001$ and $\alpha_w = 0.1$. The exploration noise that is added to the controller command is normally distributed with zero mean and variance $\sigma = 0.01$ Nm.

In the simulations with the standard delayed RL scheme, we experienced that the controller could not be trained properly. The problem is that, if the critic's surface is flat over a part of the input space, regardless of the current control performance, the temporal difference is almost zero and the controller stops adapting. This problem has been already discussed in the literature [14]. Berenji *et al.* [4] partially resolve this problem by restarting the learning mechanism when a failure occurs. Since there is no failure situation defined in our setting, another approach is proposed here.

The main idea is to use a better estimate for the predicted discounted sum when adapting the critic and the controller. When a control action is evaluated after several time steps, more immediate reinforcements (instead of one) are received, which yield a better estimate of the criterion. The outline of the procedure is as follows.



(a)



(b)

Fig. 8. The antecedent membership functions and the initial control surface of the fuzzy controller.

1) Predict the discounted sum for the expected control action and perform one step of exploration (randomly modify the control action and apply it to the process).

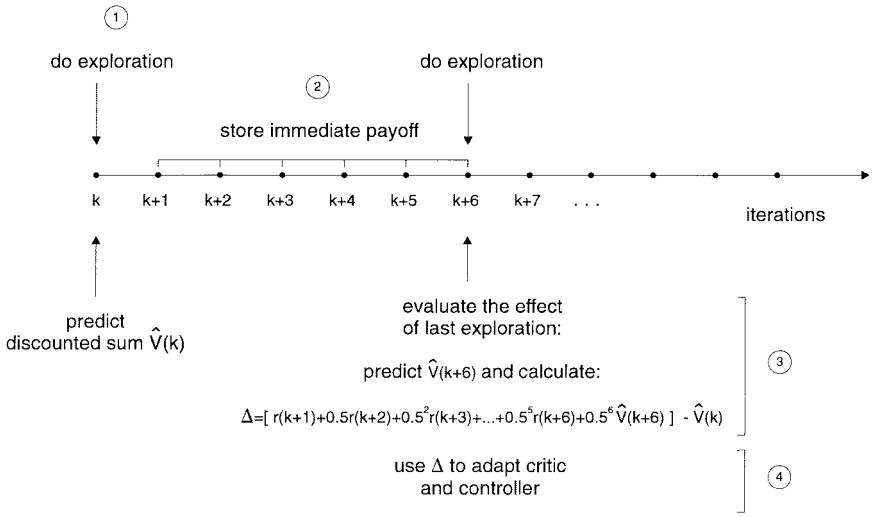


Fig. 9. The proposed alternative reinforcement learning scheme based on a more accurate estimate of the prediction criterion. The encircled numbers refer to the steps of the procedure outline given above.

2) Follow the current control policy keeping the parameters unchanged and store the immediate reinforcements.
3) After n iterations, evaluate the modified control action by estimating the discounted sum from the stored reinforcements; observing more time steps improves the estimate.
4) Use the difference between the predicted and observed criterion to adapt the critic and the controller; repeat by going to step 1).

After n time steps, the temporal difference used to adapt the critic and the controller can be written as

$$\Delta = [r(k+1) + \gamma r(k+2) + \gamma^2 r(k+3) + \cdots + \gamma^{n-1} r(k+n) + \gamma^n \hat{V}(k+n)] - \hat{V}(k). \quad (10)$$

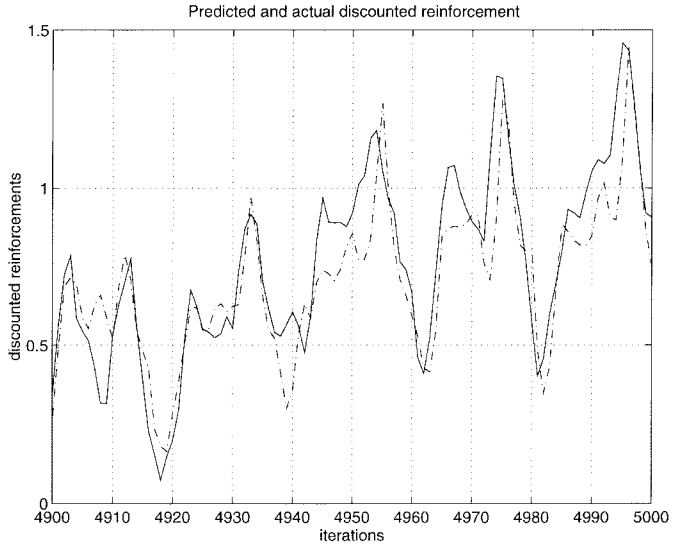
Note that the term between square brackets now represents a more accurate estimate of the criterion because it is mainly determined by the n received immediate reinforcements. The number of steps n is process dependent and should take into account the dynamics and time delays in the system. The update laws for the critic and the controller are given by (7) and (9), respectively, but the temporal difference (TD) error is replaced by (10). In this learning scheme, the adaptation stops when the maximum of the critic surface is reached. A disadvantage of this method is a longer learning time since adaptation is only performed after each n iterations.

For the satellite control, we have chosen the following setup for the proposed learning scheme (see Fig. 9). The number of iterations before an evaluation is performed is chosen to be $n = 6$. This choice is based on the maximum process delay, which is about 3 s (one iteration comprises 0.5 s). Consequently, the effect of a control action should be noticeable within this period. A range of discount factors were tested and the value $\gamma = 0.5$ resulted in an acceptable prediction error. Note that with these parameters, the contribution of the prediction at the evaluation time step $\gamma^n \hat{V}(k+n)$ is rather small and is therefore neglected. Consequently, the estimate of the criterion [term between square brackets in (10)] is completely determined by the immediate payoffs.

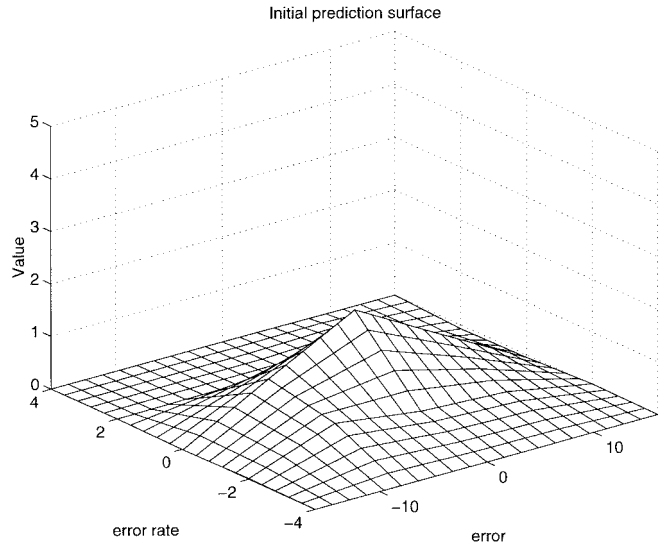
The procedure is the following: first the critic is trained off-line with the discount factor $\gamma = 0.5$ and the learning rate $\alpha_w = 0.1$. In Fig. 10(a), the prediction of the discounted sum of the immediate reinforcements is plotted. It is assumed that the prediction error of the critic is mainly caused by the stochastic part of the system (noise). In Fig. 10(b), the prediction surface after 5000 learning iterations is given. It appears that the criterion has one optimum at zero. Recall that the criterion is maximized, because the immediate reinforcement represents a reward. After 5000 learning iterations, the adaptation of the controller (with the learning rate $\alpha_v = 0.001$) is also started by adding the exploration noise. The number of iterations is 10 000, but since $n = 6$, this period actually consists of 1666 adaptation steps.

In Fig. 11(a), the trajectories of the consequent parameters of the controller during the adaptation are plotted. The parameters converge, but due to the constant exploration, the algorithm keeps searching for better actions, which results in small fluctuations around the determined parameter values. The obtained controller surface is plotted in Fig. 11(b). The adapted rule base of the controller is given in Table II. When the error rate is close to zero, the adapted controller takes less action for different errors than the initial rule base, given in Table I. The controller takes more action when the error and the error rate are of opposite signs and small. Also more action is taken when the error is around zero and the error rate is negative or positive small. This *a posteriori* analysis of the adapted control law indicates that the changes in the controller input—output mapping qualitatively agree with what one would expect the controller to do in order to satisfy the goals defined in Section IV-B.

To illustrate the performance of the adapted controller, a simulation run of 1000 s of the satellite with the new controller is performed. Fig. 12(a) and (b) shows the improvement of the attitude control in the time domain. From the plots, one can see that the attitude error and its rate are significantly reduced. The rms attitude error decreased from the initial value 3.43 to 0.77 arcseconds The rms error of the attitude rate decreased

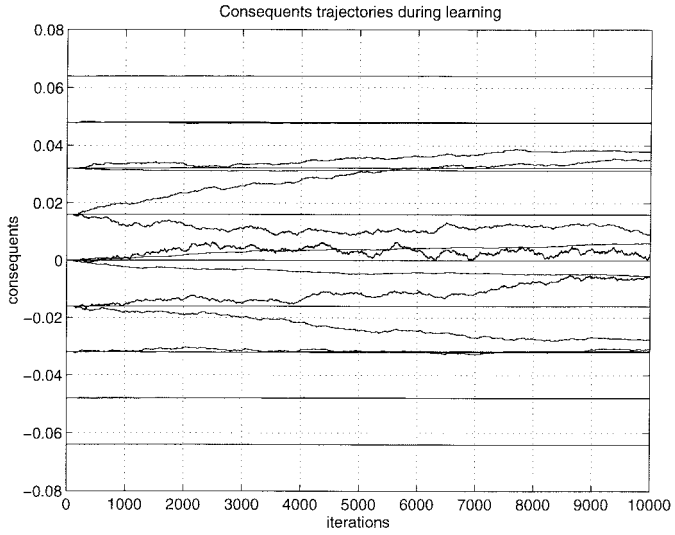


(a)

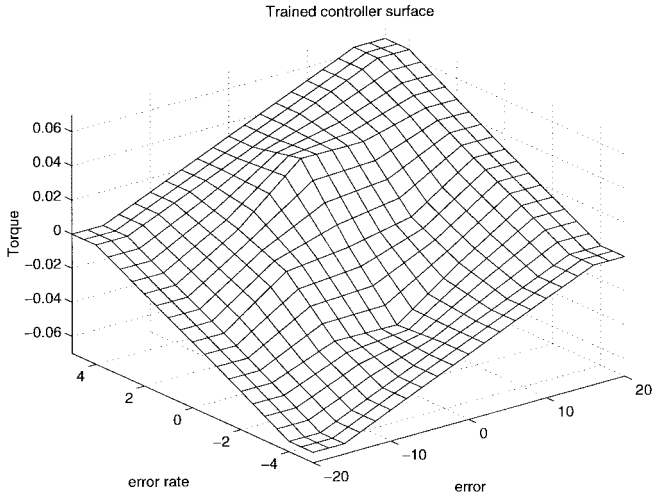


(b)

Fig. 10. Performance of the critic after off-line training.



(a)



(b)

Fig. 11. Fuzzy controller after adaptation.

TABLE I
INITIAL TORQUE VALUES OF THE FUZZY CONTROLLER (10^{-2} Nm)

	error rate				
error	NB	NS	ZE	PS	PB
NB	-6.400	-4.800	-3.200	-1.600	0
NS	-4.800	-3.200	-1.600	0	1.600
ZE	-3.200	-1.600	0	1.600	3.200
PS	-1.600	0	1.600	3.200	4.800
PB	0	1.600	3.200	4.800	6.400

TABLE II
FINAL TORQUE VALUES OF THE FUZZY CONTROLLER (10^{-2} Nm)

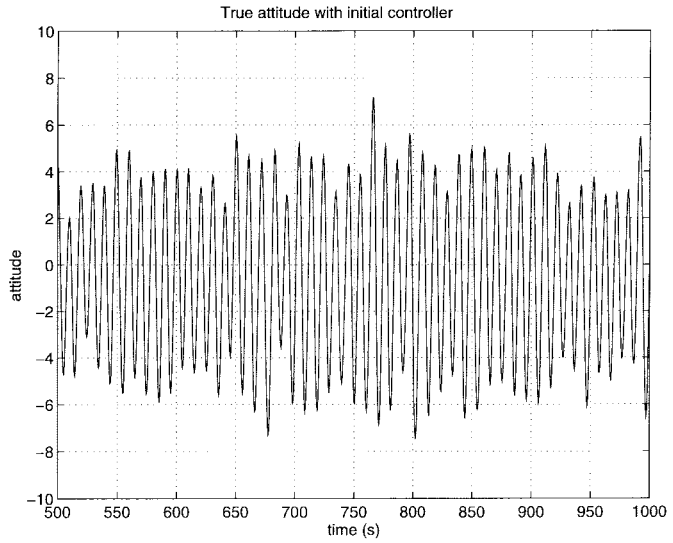
	error rate				
error	NB	NS	ZE	PS	PB
NB	-6.400	-4.780	-3.159	-1.601	0
NS	-4.796	-3.340	-0.535	0.670	1.600
ZE	-3.202	-3.395	-0.092	3.828	3.214
PS	-1.600	-0.754	-0.668	3.804	4.805
PB	0	1.597	3.136	4.781	6.400

from 2.04 to 0.30 arcseconds/s. Even though these values are clearly better than the initial performance of the linear fuzzy controller with the Butterworth filter, they are comparable with the performance of the original LQG controller with the Kalman filter (see Section II and [10]). The advantage of the RL-based approach is that the performance is not related to one specific situation and the controller can be easily adapted if the noise conditions change.

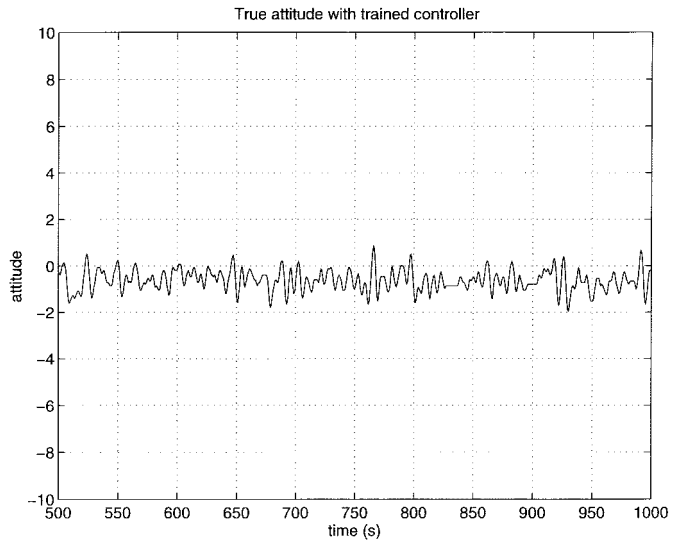
Fig. 13(a) and (b) shows the limit cycle in the phase plane before and after adaptation. From these figures, one can see that the limit cycle is not centered around zero but has a small offset. However, as stated in Section IV-B, the control goals were defined to reduce the limit cycle.

V. CONCLUSIONS

In order to reduce the limit cycle of a satellite, an adaptive fuzzy controller is applied. The results show that the nonlinear

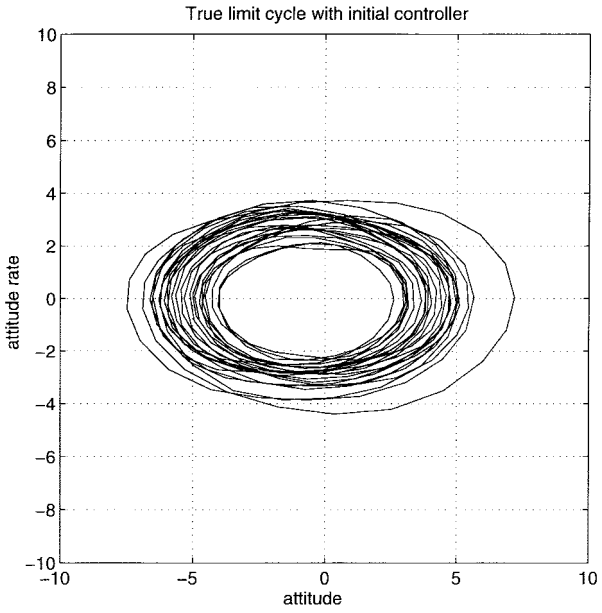


(a)

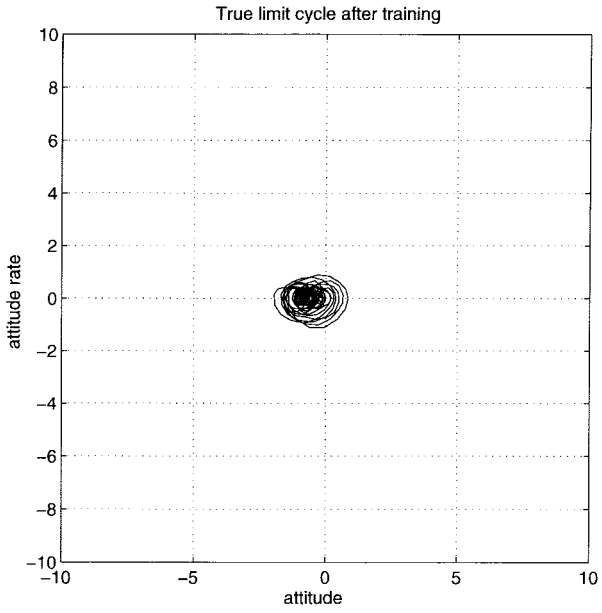


(b)

Fig. 12. Comparison of the initial and the adapted fuzzy controller in the time domain.



(a)



(b)

Fig. 13. Comparison of the initial and the adapted fuzzy controller in the phase plane.

fuzzy controller can cope with the sensor noise and nonlinearities. The controller is tuned by means of RL without using any model of the sensors and of the satellite. The main idea of RL is to find the optimal mapping between system states and controls through exploration and evaluation of possible control commands.

The reinforcement signal is computed as a fuzzy performance measure, using a noncompensatory aggregation of two control subgoals. This approach is very flexible since it allows for an accurate formulation of the control criteria and aggregation of the different subgoals. Moreover, the continuous reinforcement signal provides more detailed information about the controller performance than just binary values. Convergence of the reinforcement learning scheme is achieved by computing the temporal difference error over several time steps and adapting the critic and the controller at a lower sampling rate.

Both the critic and the controller are implemented by the TS rules with constant consequents. The TS scheme is computationally efficient, transparent to interpretation of both the initial and the adapted controller, and provides intuitively understandable and consistent rules.

ACKNOWLEDGMENT

The authors would like to thank L. Karsten of Fokker Space B.V., Leiden, The Netherlands, for providing a simulation model of the satellite and to their colleague J. Sousa for proofreading the paper.

REFERENCES

[1] Kandel and Langholz, Eds., *Fuzzy Control Systems.* Boca Raton, FL: CRC, 1994.

[2] C. T. Lin, *Neural Fuzzy Control Systems with Structure and Parameter Learning.* Singapore: World Scientific, 1994.

[3] H. R. Berenji and P. Khedar, "Learning and tuning fuzzy logic controllers through reinforcements," *IEEE Trans. Neural Networks,* vol. 3, pp. 724–740, Sept. 1992.

[4] H. R. Berenji, R. N. Lea, Y. Jani, P. S. Khedar, A. Malkani, and J. Hoblit, "Space shuttle attitude control by reinforcement learning and fuzzy logic," in *Proc. 2nd Int. Conf. Fuzzy Syst.,* San Francisco, CA, Mar. 1994, pp. 1396–1401.

[5] A. Barto, R. Sutton, and C. W. Anderson, "Neuron-like adaptive elements that can solve difficult learning control problems," *IEEE Trans. Systems, Man, Cybern.,* vol. 13, no. 5, pp. 834–846, Sept./Oct. 1983.

[6] G. Schram, L. Karsten, B. J. A. Kröse, and F. C. A. Groen, "Optimal attitude control of satellites by artificial neural networks: A pilot study," in *Proc. IFAC Symp. Artificial Intell. Real-Time Contr.,* Valencia, Spain, Oct. 1994, pp. 185–190.

[7] C. W. Anderson, "Strategy learning with multilayer connectionist representations," in *Proc. 4th Int. Workshop Machine Learning,* Irvine, CA, June 1987, pp. 103–114.

[8] R. S. Sutton, "Learning to predict by the method of temporal differences," *Machine Learning,* vol. 3, pp. 9–44, 1988.

[9] C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning," *Machine Learning,* vol. 8, pp. 279–292, 1992.

[10] W. M. Buijtenen, "Adaptive fuzzy control of satellite attitude by reinforcement learning," M.Sc. thesis, Delft Univ. Technol., Delft, The Netherlands, Dec. 1995.

[11] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its application to modeling and control," *IEEE Trans. Syst., Man, Cybern.,* vol. 15, pp. 116–132, Jan. 1985.

[12] R. E. Bellman and Lotfi A. Zadeh, "Decision-making in a fuzzy environment," *Management Sci.,* vol. 17, no. 4, pp. 141–164, 1970.

[13] D. Nauck and R. Kruse, "A fuzzy neural network learning fuzzy rules and membership functions by fuzzy error backpropagation," in *Proc. IEEE Int. Conf. Neural Networks,* San Francisco, CA, Mar. 28–Apr. 1 1993, pp. 1022–1027.

[14] W. T. C. Luenen, P. J. de Jager, J. van Amerongen, and H. M. Franken, "Limitations of adaptive critic control schemes," in *Proc. Int. Conf. Artificial Neural Networks,* Amsterdam, The Netherlands, 1993, pp. 810–813.

Walter M. van Buijtenen was born in 1969, in Breda, The Netherlands. He received the M.Sc. degree in 1995 in electrical engineering from the Control Laboratory, Delft University of Technology, Delft, The Netherlands.

Currently, he is employed at KEMA B.V., Arnhem, The Netherlands. His research interests concern the practical application of fuzzy logic and neural networks to industrial processes. He is currently working on a safety monitoring and advisory system for the Dutch power network based on fuzzy logic.

Gerard Schram received the M.Sc. degree in mechanical engineering from the Delft University of Technology, Delft, The Netherlands, in 1993. He is currently working toward the Ph.D. degree at the Department of Electrical Engineering, Delft University of Technology.

During one year, he was a Research Assistant in the Autonomous Systems Group at the Department of Computer Science of the University of Amsterdam, The Netherlands. His current research involves the application of computational intelligence to the control of aeronautical systems.

Robert Babuška was born in 1967 in Prague, Czechoslovakia. He received the M.Sc. degree in control engineering from the Czech Technical University, Prague, in 1990, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 1997.

Currently, he is an Assistant Professor at the Control Laboratory of the Electrical Engineering Department, Delft University of Technology. His main research interests include identification and control of nonlinear systems, fuzzy set theory, and fuzzy systems modeling.

Henk B. Verbruggen received the M.S. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 1963.

Since 1963, he has been a Staff Member of the Control Engineering Laboratory at the Electrical Engineering Department, Delft University of Technology. In 1976 he was appointed Associate Professor and in 1980 he was appointed a full Professor. He is author and coauthor of more than 175 publications. His research interests include adaptive and model-based predictive control, knowledge-based real-time control, fuzzy systems and neural networks for modeling and control, and supervision, planning, and scheduling of industrial plants.

Mr. Verbruggen served as Chairman of the Technical Committee for Artificial Intelligence in Real-Time Control of IFAC and Editor of the IFAC-affiliated journal "Engineering Applications of AI." He was a member of two ESPRIT working groups on Fuzzy Control (FALCON) and Model-Based Predictive Control (CIDIC), and he is a member of ESPRIT basic research project of multivariable fuzzy control (FAMIMO). He was Chairman of the Control Laboratory, Chairman of the Control Engineering Division of the Royal Dutch Institute of Engineers (NMO of IFAC), and currently he is a Vice-Dean of the Department of Electrical Engineering of TU Delft.