A Learning Model for Intelligent Agents Based on Classifier Systems and Approximate Reasoning

Jalal Baghdadchi

Department of Electrical Engineering
Alfred University
Alfred, NY 14802-1205
Email: baghdadchi@alfred.edu

Abstract

The objective of this study is to synthesize a learning model capable of successful and effective operation in hard-to-model environments. Here, we are presenting a structurally simple and functionally flexible model. The model follows the learning patterns experienced by the humans. The novelty of the adaptive model lies on the knowledge base, dual learning strategy, and flexible reasoning. The knowledge base is allowed to grow for as long as the agent lives. Learning is brought about by the interaction between two qualitatively different activities leaving long-term and short-term marks on the behavior of the agent. The agent reaches conclusions using approximate reasoning. The focus of the model, the agent, starts life with a blank knowledge base. It learns as it lives. Classifiers are used to represent individual experiences. We demonstrate the functioning of the model through a case study.

1. INTRODUCTION

The effective growth of a knowledge base is the main concern in developing a learning model. An ideal learning model envisioned for actual use should be flexible and reasonably easy to implement. While simple in structure, rigid and mathematically precise learning models are generally ineffective in expressing complex operating environments. Our daily lives and experiences suggest that a human-like learning strategy with all its flexibilities is better suited for successful functioning in a hard-to-model environment. A learning model based on the learning patterns followed by humans and relying on structurally uniform packets of information offers the desired modeling flexibility and the structural simplicity.

Human being is the perfect learning machine. At the start of his life, he begins to learn with just a few instincts to guide him. As he accumulates knowledge, he constantly strikes a balance between instincts, raw observations, and past experiences in formulating new experiences. The approximate nature of the human learning leads to a superior performance in compiling a knowledge base. This warrants a closer study of the human learning process and its possible inclusion in a machine level implementation.

2. THE MODEL

The objective of this study is to synthesize a learning model capable of successful operation in hard-to-model environments where the relation between the input and output spaces is often non-linear. Here we are presenting the model and its implementation. The structurally simple model follows the learning strategies experienced by humans. The ever-growing quality of the human knowledge, forming new experiences by following the past experiences as well as responding to fresh stimuli, and approximate reasoning in drawing conclusions are among the main characteristics of human learning. The adaptive model presented here reflects all of these characteristics.

Learning is a way of establishing the relationships between inputs and outputs of a system. It enables a learning agent to develop the capability to map the representation of a state into a probability distribution over a set of actions. Learning leads to the compiling of a knowledge structure that expresses the causal relationship between the inputs and the outputs. Humans do acquire a good deal of their lifetime knowledge through experience. Intuitively, relying on experience is the most natural way of embarking on a learning experience. Following the old saying "failure is the surest path to success", learning through experience is the least complex and most flexible method of compiling a knowledge structure (Jang, Sun, and Mizutani, 1997). Experience however, is not the sole force regulating the human behavior. Often our actions are influenced by what transpires around us. We may not always have a past experience pertaining to the situations we face. In the domain of educational psychology it is widely believed that human learning is hybrid in nature (Grossberg, 1995); that it takes place by processing both the explicit data and the contextual information. Combined processing of past experiences and current sensory information is therefore key to the formulation of a comprehensive self-learning strategy.

The model presented here is characterized by a realistic portrayal of human learning experience and its adaptation to machine level implementation. Its connectionless architecture makes it possible to express the ever-growing nature of human knowledge (Baghdadchi

2000). It involves both past information and current stimuli in forming fresh experiences. By using fuzzy logic the model effectively expresses the non-deterministic nature of learning. The focus of the model, the learning agent, starts life with a blank knowledge base. Learning takes place as a result of accumulation of experiences. The agent follows a dual functioning strategy. Learning is driven by the successes and failures, by trial and error, and by the reward system. Structurally uniform packets of information called classifiers represent experiences.

2.1. Knowledge Base

The knowledge base embodies the life-long experiences of the agent. The knowledge base is composed of classifiers (Goldberg, 1989), each representing a single unique experience. Each classifier is composed of a state-and-action pair along with the pair's measure of usefulness. We refer to these elements as *state*, *action*, and *value*. A given classifier $C_{i,j}$ and its components are denoted by:

$$C_{i,j} = \{s_i, a_j, v_{i,j}\}$$

with s_i, a_j, and $v_{i,j}$ representing state *i*, action *j* paired with state *i*, and the *value* of the pair. Note that any given state may be paired with different actions. The agent will continue to form and accumulate new classifiers so long as its mission is in progress.

2.2. Learning In Detail

The model employs a dual functioning strategy: going back in time and retrieving past experiences and evaluating the current state of the mission. The agent's past experience and its perception of the state of mission are the two parameters governing its behavior. At all times the agent is unaware of what lies ahead. Rather than blindly following past experience, it combines past experience with its current perception of the state of mission. The state of mission is characterized by a finite set of classifiers depicting the agent's most recent activity. We will refer to this set as the moving history. Unlike the knowledge base, which continues to grow as long as the agent lives, the number of state-and-action pairs in the moving history remains unchanged. At any given time, the newly received stimuli are processed in the context of the agent's knowledge and experience. The agent first matches the sensed state with the *state* element of the classifiers in the knowledge base. Based on the absence or presence of matched classifiers and their attributes, and the state of mission, the agent then formulates a response.

Learning results in formation of new experiences, which are recorded in the form of classifiers and added to the knowledge base. The *value* elements are the modifiable parameters of the knowledge base, accounting for the adaptiveness of the learning model. Learning is unsupervised. At no point in time along the way does the agent know if the prepared response will take it to the objective, nor does it have any measure of judging the prepared response against the ideal response. The action a_{next} selected as a response to state s_i is a function of s_i, past experiences resulting from actions taken departing from the state s_i, and the state of mission denoted δ:

$$a_{next} = f(s_i, C_{i,1}, C_{i,2}, \ldots, C_{i,j}, \delta)$$

δ is obtained by evaluating the temporal difference of the subset of classifiers forming the moving history. Thus, selecting a response becomes a joint probability distribution from the experience and the state of mission onto a set of actions. A flexible reasoning module determines the right combination of each parameter's participation in the preparation of response at any given time.

Once the response is prepared a new classifier is formed by pairing the sensed state and the selected action. The moving history is updated by adding the fresh state-and-action pair and deleting the oldest classifier in the set. The new classifier is also added to the knowledge base. If the mission is successful, the sequence of classifiers leading the agent to the objective receives reward. The reward is added to the *value* elements of the classifiers. The details of the reward assignment are discussed in the next section.

2.3. Reward Assignment

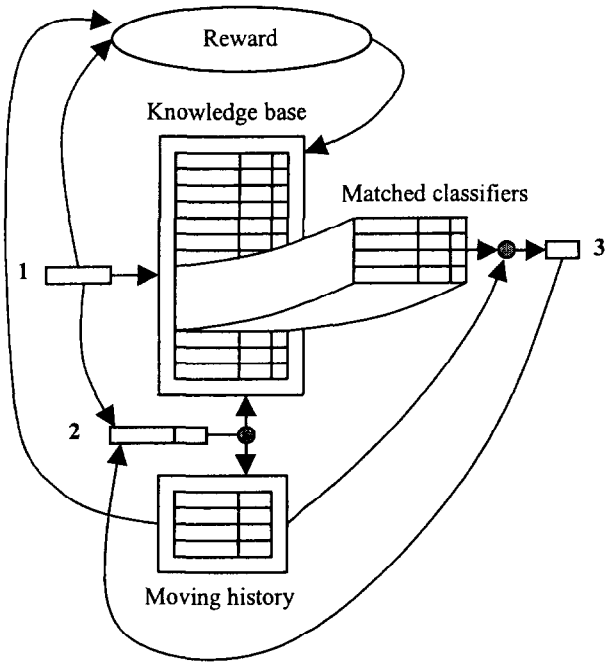
The model is adaptive in nature. The value part of the classifiers represents the measure of usefulness of each state-and-action pair, and the strength with which it influences the agent's future actions. Sequences of classifiers leading the agent to a desirable situation (i.e., accomplishment of the mission) receive reward. The reinforcement signal administered in the form of a reward sets the recipient classifiers apart from the rest. It also reflects the potential usefulness of the recorded experience. Once the end game is in sight, the efficacy with which the classifier sequence has lead to the final objective becomes clear. The strength of the aggregate reward and the mode of distribution may be adjusted according to the efficacy level of the sequence. The last classifier of the sequence receives the highest reward. Going back in time from the achievement of the final objective, the strength of the reward decreases. In the event of a success, in practice only a handful of classifiers may be assigned a reward. The application of the reinforcement signal to classifiers going *k* steps back in time obeys:

$$v_k = v_{k=0} e^{-f(k)}, \; k = 0,1,2,\ldots$$

$$\sum_k v_k = v_{aggregate},$$

Where v_k indicates the reward at step *k*, $v_{k=0}$ the maximum reward, and $v_{aggregate}$ the total reward, and *f(k)* has a positive value. The above relations mean that the adjustments to the *value* of the active classifiers going *k* steps back in time are exponentially weighted, with the more recent classifiers receiving the higher rewards (Sutton, 1990). The block diagram in Figure 1 depicts the

functioning of the learning model.



1: Sensed state
2: Newly formed state-and-action pair
3: Selected action

Figure 1. Functioning of the Learning Model

3. CASE STUDY

In this section we introduce the application of the learning model to a learning problem. The implementation of the chosen case reflects all the characteristics of the learning model. The ever-growing quality of the knowledge base, the dual nature of the learning strategy, adaptiveness of the model, and the approximate nature of learning are all demonstrated in this case study. Here, we first introduce the problem, then outline the approach, explain the main components, and present the implementation details.

3.1. The Problem

Consider a two-dimensional bounded area. The area is divided into blocks, some of which are occupied by obstacles. The area is also bounded by obstacles along its sides. The learning agent roams within the area with the objective of capturing the target (Wilson, 1985). Through trial and error and rewarding of successful experiences the agent is expected to develop efficient ways of getting to the target from any point on the terrain. The knowledge base is allowed to grow continuously during the life of the agent. The dual nature of learning is represented by the interaction of experience and the agent's perception of the state of mission. Experience is compiled gradually. It evolves continuously and participates in guiding the agent throughout its life. The agent's perception of the state of mission changes from one scenario to another. Its effects on the behavior of the agent are temporary. Successful learning requires that the two parameters participate in decision making with varying degrees at different stages of the agent's life (Kosko, 1993). A flexible reasoning module determines the appropriate combination of each parameter's participation at any given time.

3.2. The Operation

The agent can move one block at a time, in a total of eight possible ways: N, NE, E, SE, S, SW, W, and NW. It cannot move into a block containing an obstacle. The agent starts life with a blank knowledge base. The learning scheme thrives on recording and recalling experiences. Each single move is represented by a classifier. The collection of classifiers forms the knowledge base. Each classifier is composed of a state, a corresponding action, and a value representing the pair's measure of usefulness. Note that any state vector may be paired with different actions. The state represents the eight blocks immediately adjacent to the agent. Each block of the sensing area may be blank (unoccupied), occupied by an obstacle, or occupied by the target. The action part is simply the address of the block that the agent is suggested to move into next, relative to the current position. The combination of the state and the action forms an IF-THEN structure, which along with value forms a standard information packet summarizing a single move. The recorded experiences are used every time the problem is attempted. As more and more attempts from different parts of the terrain end in success, the agent constructs more efficient ways of reaching the target.

The image of the agent's immediate surroundings forms the current state. Prior to each move, the agent senses its surroundings and forms an image of them. It then searches the collection of its past experiences, for possible matches with the current state. The direction of the agent's last move defines its perception of the state of mission. We will call this the localized sense of direction. At any given time the next move is selected based on the current state, past experiences involving the current state, and the localized sense of direction. The two parameters influencing the selection of the next move, past experience and the localized sense of direction are resolved using fuzzy logic. Classifier sequences that guide the agent to the target receive rewards. The accumulated rewards of each classifier represent its respective influence in conducting the future moves of the agent. The level of reward depends on the efficiency of the sequence. The reward is distributed between the last three to five active classifiers. There is no negative reward. For the most efficient case, the five classifiers leading to the target receive rewards. The action a_0 of classifier C_0 in the sequence, which brings the agent to the immediate vicinity of the target, receives the highest reward. The distribution of reward can be expressed approximately by:

$$C_{k-value} = Ae^{-0.57k}, k \in [0,4]$$

where $C_{k-value}$ is the value part of classifier C_k, k is the step index going back in time, and A is the maximum reward.

3.3. Selection of The Next Move

The selection module carries out the selection of next move. The inputs to the selection module are the past experience and the localized sense of direction. The output is the address of the block that the agent next moves into, relative to its current position. The past experience is represented by a subset of knowledge base. The subset is composed of the classifiers with states identical to the current state. The localized sense of direction is simply the direction of last move. The first step in the selection process is to compare the current state with classifiers in the knowledge base. If a match is not found and an obstacle does not occupy the block along the previous direction, the agent moves into that block. If the block is occupied, the agent takes a path with the least deviation from the previous direction. If matched classifiers are found, the resultant of the matched classifiers (based on the location and strength) is determined. The resultant is expressed by the cumulative strength of the past experiences and the angle it makes with the direction of previous move. All angles are presented in units of blocks:1 block = $\frac{\pi}{4}$. Positive angles are defined in clockwise direction. Fuzzy logic (Zadeh, 1965) (Zimmermann, 1996) is used to combine the effects of the two parameters. The inputs to the module are the cumulative strengths of the relevant past experiences and the angle it makes with the direction of the previous move. The output of the module is the address of the block the agent is suggested to move into next, relative to its current location. The input and output membership functions are shown in Figure 2. The first input to the fuzzy module, ***experience***, is described by the three trapezoidal membership functions ***weak***, ***strong***, and ***dominant***. The second input, ***angle***, is described by seven trapezoidal membership functions ranging from ***opposite_negative*** to ***zero*** and ***opposite_positive***. The output variable ***next_move*** is represented by seven triangular membership functions ranging from ***far_left*** through ***straight*** to ***far_right***. Both inputs are present in the antecedent part of all rules. A total of 21 rules span the entire input space.

As expected, initially the movements of the agent are primarily driven by the localized sense of direction. The localized sense of direction advises the agent to follow the direction of the move, where possible. This is analogous to an instinct calling on the agent to maintain a direct path. As more runs are attempted, the collected experience plays an increasingly more important role in determining the course of action. The gradual shift of the significance from instinct to experience is strikingly similar to the shift in behavioral patterns of intelligent biological entities.

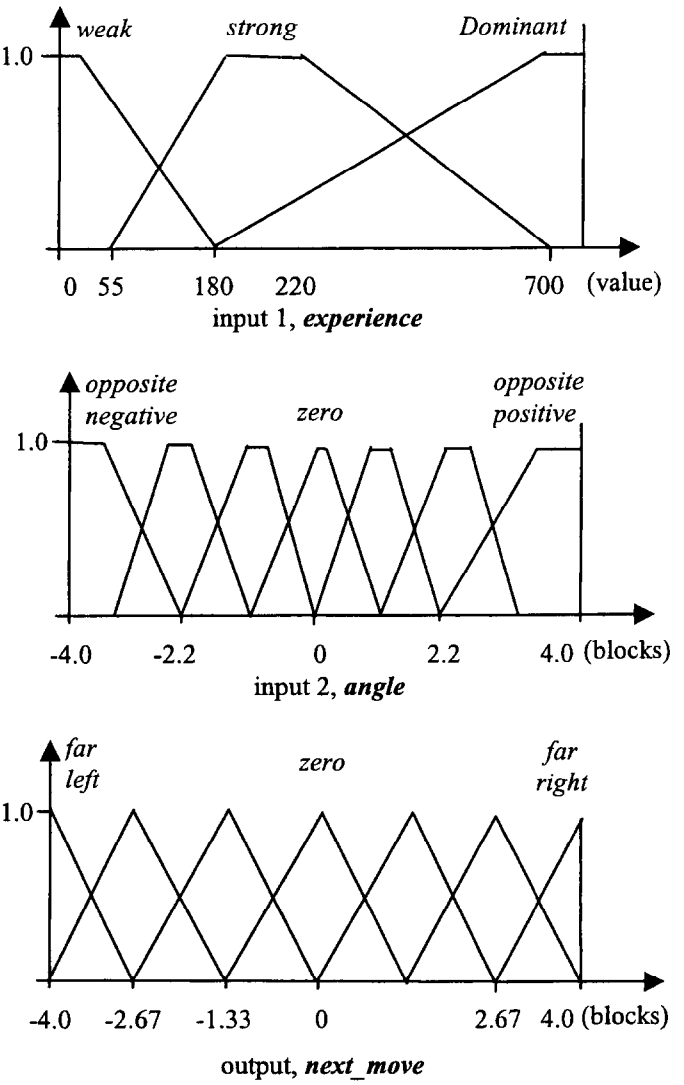


Figure 2. Membership Functions

4. SIMULATION AND RESULTS

In this section results of series of simulations are shown as the main features of learning model are incorporated in the procedure. These features are the ever-growing quality of the knowledge base, the dual learning strategy, adaptiveness of the learning model, and the approximate nature of learning. Through the results of these cases we demonstrate the impact of these features on improving the agent's performance. In all studied cases the knowledge base is allowed to grow for as long as a session of consecutive runs is in progress. The overall performance of the agent is then a function of the learning parameters and the decision-making strategy. The learning parameters are the experience and the localized sense of direction. The reward assignment determines the degree to which experience influences the behavior of the agent. This influence indicates the model's measure of adaptiveness. The behavior of the agent is also influenced by the strength with which the localized sense of direction is enforced. We introduce each of the adaptiveness and the localized sense of direction with three different strengths,

resulting in a total of nine cases (Table 1). A flexible inference strategy is introduced in case 10.

Table 1

Localized Sense of Direction	Adaptiveness		
	Minimally Adaptive	Partially Adaptive	Fully Adaptive
Not Enforced	case1x	case1y	case1z
Partially Enforced	case2x	case2y	case2z
Fully Enforced	case3x	case3y	case3z

The strength of adaptiveness basically depends on the extent to which the knowledge base adapts to a particular experience, which in turn depends on the rewards it receives. In the event of a capture, the last few classifiers receive rewards. In minimally adaptive cases, only the classifier that brings the agent to the immediate proximity of the target receives reward. For partially adaptive cases the last one, two, or three classifiers receive rewards. For fully adaptive cases the last five classifiers receive rewards. The localized sense of direction is implemented by encouraging the agent to maintain a straight path where possible. In full strength, unless the agent comes face-to-face with a wall, it is not allowed to return to the vicinity of its previous position. The 3-block-no-return policy is not enforced if the agent does come face-to-face with a wall. Mid-strength localized sense of direction, employs a 1-block no-return policy. For zero strength the localized sense of direction is not enforced. All cases are tried for 1000 blocks of 200 consecutive runs. An attempt where the agent does not reach the target by the 120^{th} move is considered a failure. In all performance graphs, the values on the horizontal axis represent the consecutive run indices, 1 to 200. The values on the vertical axis represent the average number of steps required for capture. The results of simulation for the fuzzy case are shown in Figure 3.

With localized sense of direction not enforced, for the minimally adaptive case, in the event of success, reward is assigned to the classifier, which brings the agent to the immediate proximity of the target. In this case the selection of next move is based on experience only. During the first few runs of the 200 consecutive run block due to low probability of encountering pertinent experiences in the knowledge base, the selection of next move is primarily random. Random selection rarely leads to success. In fact, the probability of reaching a target as near as just 3 blocks away, with the minimum number of steps, is quite low:

$$p_{\text{straight-to-a\_target-3\_blocks\_away}} = \left(\frac{1}{8}\right)^3 (7.5)\left(\frac{8-p}{8}\right)^3 \approx 0.8\%,$$

where p indicates the average number of occupied blocks per the 8-block sensing area and has a value of 1.49 for the simulation terrain. For the fully adaptive case, in the event of success, the last five classifiers leading the agent to the target are eligible for reward. The assignment of reward, however, depends on the efficiency of the route taken. The performance continues to improve throughout the run-block. With localized sense of direction held constant, strengthening of adaptiveness leads to gradual improvement in performance, through enhancing the learning capability (Figure 4). The presence of the one-block-no-return policy with partially enforced sense of direction, induces a measure of direction into the movements of the agent increasing its chances of getting to the target. Strengthening localized sense of direction while holding adaptiveness constant leads to scaling of the learning graph. This is clearly shown in Figure 5.

In case 10, the fuzzy case, the selection of next move is decided through the fuzzy module. While the strength of the localized sense of direction is constant throughout a block of 200 runs, the experience assumes a more vigorous role as the runs progress. Therefore, unlike the nine cases shown and discussed previously, where the ratio of the influence of the localized sense of direction and adaptiveness on performance remains constant for the duration of the run-block, here the ratio changes in favor of adaptiveness as experience becomes more pertinent.

Throughout the trials the knowledge base is allowed to grow for as long as a run session is in progress, accounting for the ever-growing quality of the agent's knowledge. There is little change in the number of moves to capture as the session progresses when only the last classifier of a successful sequence receives reward. All significant learning takes place at the early youth. Graphs in series 3 suggest that a stronger adaptive capability leads to a longer learning life and an overall better performance (Figure 4). Also, graphs in series z reveal the scaling effect of the localized sense of direction (Figure 5).

5. CONCLUSIONS

In this study we introduced a new learning model capable of successful and effective operation in complex and hard-to-model environments. The focus of the model, the learning agent, starts life with a blank knowledge base. It learns as it lives. The agent is allowed to store the knowledge for as long as it lives, which it does by forming uniform packets of information. The behavior of the agent is not solely driven by the past experience. What transpires around the agent in conjunction with its sense of mission also influences its decisions. A reward system distinguishes the useful experiences from the rest.

Through a series of trials we study the impact of both of experience and the localized sense of direction on the performance of the agent. By holding one parameter constant while varying another, we develop a qualitative understanding of the impact of each parameter on the

overall performance. We note that adaptiveness directly affects learning and indirectly affects decision making, whereas the localized sense of direction directly affects decision making and indirectly affects learning. Simulations of cases with desirable results support the idea that at the start of life external excitation in form of localized sense of direction play a dominant role in shaping the behavior of the agent. As life progresses, however, experience evoked through active means becomes the primary factor influencing the selection of next move. Here, we demonstrated the merits of a learning model constructed on the basis of general characteristics of human learning. The special human-like features of the learning model presented here, in particular the ever growing nature of knowledge base, dual learning activities, and the flexible inference, afford the agent the capability to rapidly adjust its behavior to the attributes of a highly nonlinear problem.

REFERENCES

1. Baghdadchi, J., "A Novel Learning Model for Intelligent Agents", Proceedings of the IEEE Systems, Man, and Cybernetics, SMC'00, Oct. 2000.

2. Goldberg, D. E., Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989, Chapter 6.

3. Grossberg, S., "The Attentive Brain", American Scientist, 1995, Volume 83, pp. 438-449.

4. Jang, J. –S. R., Sun, C.-T., Mizutani, E., Neuro Fuzzy and Soft Computing, Prentice Hall, Upper Saddle River, NJ,1997.

5. Kosko, B., Fuzzy Engineering, Prentice Hall, Upper Saddle River, New Jersey, 1993.

6. Mamdani, E. H., "Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis," IEEE Transactions on Computers, vol. C-26, no. 12, 1182-1191, December 1977.

7. Sutton, R. S., "Reinforcement Learning Architecture for Animats", The Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats", 1990.

8. Wilson, S. W., "knowledge Growth in an Artificial Animal", Proceedings of International Conference on Genetic Algorithms, pp. 16-23, July1985.

9. Zadeh, L. A., "Fuzzy Sets", Information and Control, vol. 8, 338-353, 1965.

10. Zimmermann, H. J., Fuzzy Set Theory and its Applications, Kluwer Academic Publishers, Boston, MA, 1996.

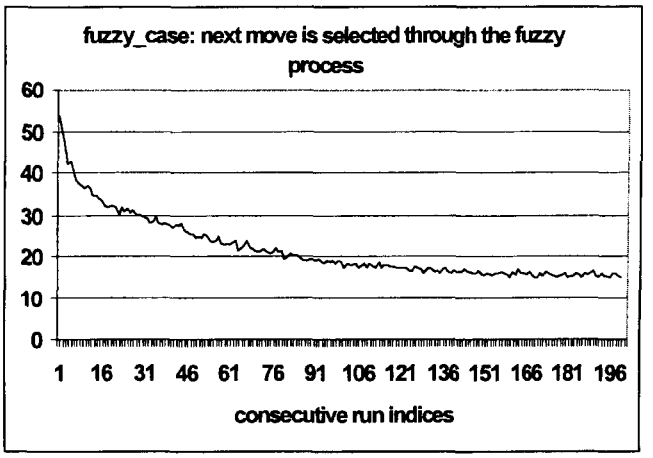


Figure 3: The fuzzy case

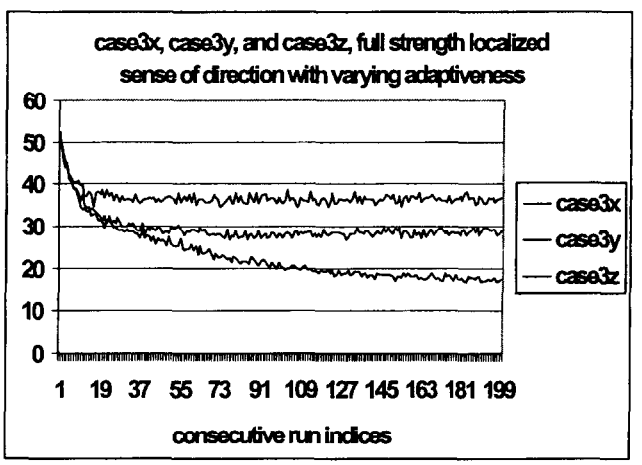


Figure 4: Effects of adaptiveness on learning

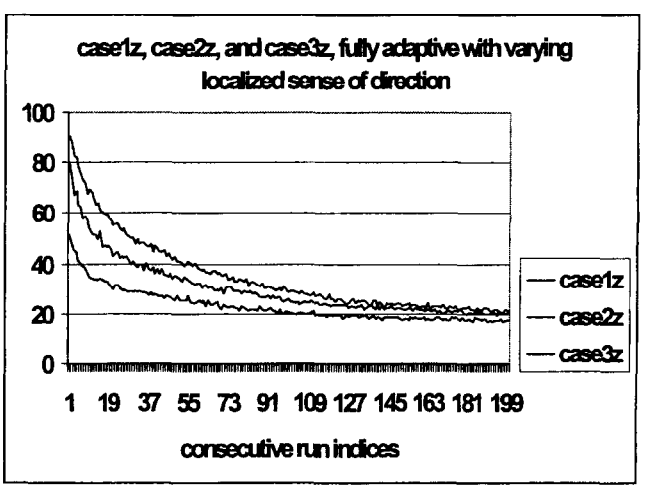


Figure 5: The scaling effect of localized sense of direction on performance