Operant Conditioning Learning Model based on BP Network

HUANG Jing[1,2], RUAN Xiaogang[1], LI Lei[2], WEI Ruoyan[1], FAN Qingwu[2], WU Xuan[1]

1. Institute of Artificial Intelligence and Robotics, Beijing University of Technology, Beijing 100124,China
E-mail: aiandrobot@163.com

2. Pilot College, Beijing University of Technology, Beijing 101101, China
E-mail: mymailhj@sohu.com

Abstract: The naissance of cognitive robotics marks that psychology is more and more highly involved in the artificial intelligence research. Inspired by psychology and ethology, we propose an operant conditioning learning model based on BP (back-propagation) network named *OCLMBP* on the basis of Skinner's relevant theory. The model is applied to the problem of obstacle avoidance with a wheeled robot. The robot controlled by the model can learn to avoid obstacles through a learning-by-doing style without any external supervision, but by the proximity sensors information as positive or negative reinforcement signals. The results are compared with original *OCLM* (operant conditioning learning model), and the proposed model has better performance.

Key Words: Operant conditioning, BP network, obstacle avoidance, cognitive robotics

1 Introduction

It is the common theme of cognitive science, artificial intelligence, and robotics to explore cognitive mechanisms and comprehend cognitive behaviors. As a result, the concept of *Cognitive Robotics* emerged. In 1997, Brooks came up with the notion *Cognitive Robotics*[1]. He explained the new idea in seven issues, such as self adaptation, development, motivation, etc. Later, more researchers joined the discussion, which indicates *Cognitive Robotics* as a new field has been found and attracted more and more interests[2,3].

One evident feature about the research of cognitive robotics is that cognitive science, especially cognitive psychology, is highly involved. Inspired by the memory theory, Stachowicz et. al[4] presented an approach to provide a cognitive robot with a long-term memory of experiences, a memory which was then testified to be able to run efficiently in robot applications involving several hours of experience. Barto et. al[5] presented a cognitive computational model and got initial results aimed at allowing artificial agents to construct and extend hierarchies of reusable skills that are needed for competent autonomy. The core of the model is just based on the intrinsically motivated learning which comes from psychology theory.

Our work is just influenced by the research approach. The inspiration comes from Skinner's operant conditioning theory[6], which was first proposed in his famous book *The Behavior of Organisms: an experimental analysis* in 1938. Skinner divided stimuli into positive reinforcement which can increase the probability of the action producing the stimulus, and negative reinforcement which can decrease the probability of the action. In another word, actions produce stimuli, then stimuli influence the probability of actions as feedback, and the impacts of different stimuli on actions are different.

Skinner's operant conditioning theory reflects the self-adaptation of agents to the environment and attracted many researchers. In 1995, Zalama et. al[7] presented an approach based on Grossberg's conditioning model to resolve the problem of obstacle avoidance with a wheeled mobile robot. The model mainly referenced the classical conditioning, in which the distance acted as conditioned stimuli(CS) while collisions act as unconditioned simuli(UCS). By experiencing a series of movements in a cluttered environment, the robot learned to avoid sensor activation patterns that predict collisions, and thereby learned to avoid obstacles. Then, Gaudiano et. al[8,9] continued to develop the model by combining with artificial neural network. The model was applied to real robots Pioneer 1 and Khepera to resolve obstacle avoidance and was proved to be efficient. Although their work was successful, their models were mainly based on the classical conditioning theory and hardly discussed the operant conditioning theory.

In 2004, Ishii et.al[10] repeated the Skinner rat experiment between rats and robots(WM-6) to investigate animal-robot interactions. Referencing to Skinner's operant conditioning, they developed a robot behavior generation algorithm that enabled the robot to autonomously show rats how to obtain food by pushing the levers. In 2005, Itoh et.al[11] presented a behavior model for humanoid robots based on operant conditioning. They implemented this behavior model into the Emotion Expression Humanoid Robot WE-4RII (Waseda Eye No.4 Refined II) and confirmed that the robot with the behavior model could autonomously select suitable behavior for the situation within a predefined behavior list. In 2007, Taniguchi et.al[12] presented a novel integrative learning architecture based on a reinforcement learning schemata model (RLSM) with a spike timing-dependent plasticity (STDP) network. This architecture models operant conditioning with discriminative stimuli in an autonomous agent engaged in multiple reinforcement learning tasks.

*This work is supported by National Natural Science Foundation of China (No.61075110; No.61375086); Key Project (No.KZ201210005001) of S&T Plan of Beijing Municipal Commission of Education; National Basic Research Program of China(973 Program) (2012CB720000); Specialized Research Fund for the Doctoral Program of Higher Education (No.20101103110007); Beijing Universities Young Talent Plan(No.YETP1610).

Despite the various study related to operant conditioning done by the researchers above, they did not present a general and formal description about operant conditioning.

Cai et.al[13,14] has done a lot of thorough work in this respect since 2010. They proposed a formal learning model based on automata theory to describe operant conditioning theory, which was applied to repeating Skinner's classic animal experiment and resolving the self-balance problem of wheeled robots. However, the application of their model was quite limited to the balance control for robots and some shortcomings existed in the model, e.g. the orientation function mentioned in the model wasn't consistent with the corresponding biological concept, which made the model not perfect enough in theory. Huang et.al[15,16] therefore revised the model, proposed a new one named *OCLM* (operant conditioning learning model) and proved its convergence.

Starting from the work on *OCLM*, we combine it with BP (Back Propagation) network to improve it. We apply the model to solving obstacle avoidance problems and compare the results of two models (with BP network or without it). The results show that *OCLM* based on BP network can fasten the speed of convergence and helps enhance the self-organization degree and intelligence of robots.

2 Operant Conditioning Learning Model

OCLM (operant conditioning learning model) was first proposed in the paper Operant Conditioning Learning Model in the Bionic Experiment[15]. Here we revise it slightly to combine it with BP network.

The Operant Conditioning Learning Model (*OCLM*) is defined by eight elements: $OCLM =< t, S, A, P, f, \varepsilon, \delta, L >$,and it is defined as follows:

① *t*:time parameter of the *OCLM*, and also represents the model iterative times, $t = \{t_i \mid i = 0,1, \cdots n_t\}$, t_0 represents initial establishment time of the model.

② *s*: the state space of the *OCLM*, $S = \{s_i \mid i = 1, \cdots n_S\}$, s_i represents the *i-th* state in the state space, and n_S is the size of the state space which means the number of states.

③ *A*: the set of actions in the *OCLM*, $A = \{a_k \mid k = 1, 2, \cdots n_a\}$, a_k represents the *k-th* action and n_a is the numbers of actions.

④ *P*: the probability distribution of *OCLM*, $P: S \times A \to P = \{p_{ik} \mid i = 1,2,...,n_S, k = 1,2,...,n_a\}$, in which $p_{ik} = p(a_k \mid s_i)$ represents the probability of choosing $a_k \in A$ when the agent is in the state $s_i \in S$.

P can also be written as the vector format $P = \{P_1, P_2, ..., P_{n_S}\}$, in which $P_i = \{p_{i1}, p_{i2}, \cdots p_{ir}\}$ means the probability vector for the *i-th* state and *r* means the number of possible actions at that state.

Obviously, $\sum_{k=1}^{r} p_{ik} = 1$ and $0 \le p_{ik} \le 1$.

⑤ *f*: the state transfer function in *OCLM*, $f: S \times A\big|_P \to S$,means that the state $s_i \in S$ will be transferred to $s_j \in S$ after the agent chooses the $a_k \in A$ with the probability of $p_{ik} \in P$.

⑥ ε: the negative ideal degree(*NID* for short) of each state. The *NID* is presented in this paper to calculate the orientation function. The definition is as follows:

NID is denoted as $\varepsilon = \varepsilon(S) = \{\varepsilon(s_i) \mid i = 1, 2, \cdots n_S\} \in R$ and it indicates how negative ideal the state is. The bigger value of *NID*, the farther away from ideal state for state s_i.

⑦ δ : the orientation function in the *OCLM*. The orientation function $\delta = \delta(S, A) = \{\delta_{ik} \mid i = 1, 2, \cdots n_S; k = 1, 2, \cdots n_a\}$ is introduced to simulate the biological orientation in nature. Symbol δ_{ik} means the performance changes of the system after choosing $a_k \in A$ at state $s_i \in S$. To keep identical with the biological orientation in meaning, if $\delta > 0$, the orientation is positive, which indicates that the performance of the system tends to become better, while if $\delta < 0$, the orientation is negative, which indicates that the performance tends to become worse. If $\delta = 0$, it indicates that the performance has no change. Based on the definition of *NID*, the orientation function δ is defined as follows:

Suppose action a_k is executed in the state s_i and then the state is transferred to s_j, δ_{ik} is defined as follows:

$$\delta_{ik} = \delta(\Delta\varepsilon_{ij}) \begin{cases} >0, & \Delta\varepsilon_{ij} < 0 \\ = 0, & \Delta\varepsilon_{ij} = 0 \\ <0, & \Delta\varepsilon_{ij} > 0 \end{cases} \qquad (1)$$

in which $\Delta\varepsilon_{ij} = \varepsilon(s_j) - \varepsilon(s_i)$. The orientation function is continuous in the defined interval and is monotone decreasing function about $\Delta\varepsilon_{ij}$ and its absolute value is monotone increasing about the absolute value of $\Delta\varepsilon_{ij}$.

Here we provide the explanation for the definition as follows. If $\Delta\varepsilon_{ij} > 0$, it means the negative ideal degree is increasing, i.e. the system performance becomes worse, therefore the orientation function $\delta < 0$. Moreover, the bigger $\Delta\varepsilon_{ij}$, the smaller δ. On the contrary, if $\Delta\varepsilon_{ij} < 0$, the negative ideal degree is decreasing, which indicates the system performance becomes better, so the orientation function $\delta > 0$. Similarly, the bigger $\Delta\varepsilon_{ij}$, the smaller δ. Finally, if $\Delta\varepsilon_{ij} = 0$, the negative ideal degree has no change, indicating the system performance doesn't change, so the orientation function $\delta = 0$.

⑧ *L*: the learning mechanism of the *OCLM*, $L: P(t) \to P(t+1)$ ($P(t)$ is the probability distribution at time *t*). *L* will adjust the probability distribution according to operant conditioning, which means that the possibility of the actions causing the positive reinforcement will be increased

while the probability of the actions causing the negative reinforcement will be decreased.

In summary, the working principle of the *OCLM* can be described as follows: suppose at time *t*, the negative ideal degree of the state $s_i \in S$ is ε_i and $a_k \in A$ is selected with the probability $p_{ik} \in P$. Then the state will be transferred to state s_j. After receiving the stimulus from the environment for this action, the negative ideal degree is changed to ε_j. Then the orientation function $\delta(\Delta\varepsilon) = \delta(\varepsilon_j - \varepsilon_i)$ can be calculated. Based on it the probability distribution function *P* can be adjusted according to the learning mechanism *L*. After multiple rounds of iterative calculation, the model can acquire the optimal action and form operant conditioning.

3 OCLM based on BP network

In 1986, Rumelhart el.al[17] described in Nature a new learning procedure, back-propagation, for networks of neuron-like units. Ever since then the BP algorithm was attracting more and more attention. Since it has been proved to have strong representation capability[18,19], BP is the most widely used algorithm in applications. So, we choose BP network to realize *OCLM*.

The architecture of *OCLMBP*(Operant Conditioning Learning Model based on BP network) is shown in figure 1. It consists of two parts: a 3-layer BP network and one operant conditioning computing module. BP network realizes such a mapping from state space (sensorial information, which is corresponding to *S* in *OCLM*) to action set and its probability distribution(motorial information, which is corresponding to *A* and *P* in *OCLM*) while the computing module takes operant conditioning theory as complementary principle to adjust the weight of the network in order to find the best match action under a certain state.

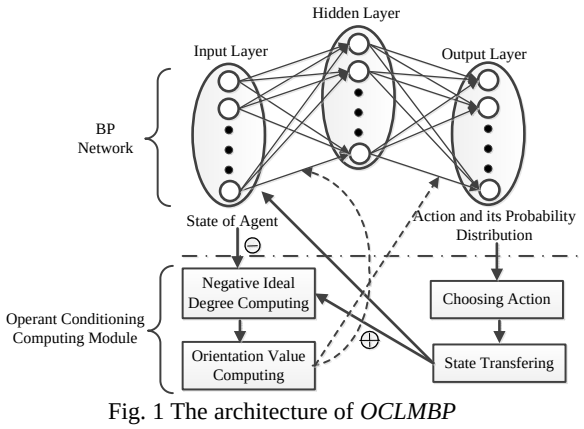

Fig. 1 The architecture of *OCLMBP*

The arrow-line in the figure represents the direction of data flow from one unit to another. The arrow-line starting from orientation value computing unit is broken to show that the orientation value isn't the only one involved in the adjustment of network weight. It's just complementary to the traditional way of back-propagation algorithm.

4 Simulation Experiment Design and Results

Robots with *OCLMBP* just like animals or human-being can autonomously learn skills based on operant conditioning principle through interacting with the environment. Furthermore, *OCLMBP* can take advantage of BP network to get better performance than *OCLM*. To prove the conclusion, we simulate an environment with a number of obstacles in which a mobile robot with sonar sensors must avoid them during navigation. The whole experiment includes 2 parts: 1. the robot navigates in the environment with fixed obstacles, 2. the robot navigates in the environment with variable obstacles whose position is set randomly and change every time when the experiment runs. The controller of the robot is the computational implementation of *OCLM* or *OCLMBP*. We tested both models in the same environment and compare their ability to avoid the obstacles, which indicates the self-learning ability of the model.

4.1 Description of the environment and the robot

The simulation environment is a 4m square with 10 obstacles. In the fixed position experiment, the obstacles scatter along the line from the start point to the end point (see fig.2). In the other case, they scatter randomly in the environment.

The simulated agent is a circular mobile robot whose radius is 0.2m. 6 sonar sensors scatter evenly around it. The measurement range of each sensor is from 0.15cm to 8m which completely covers the environment. The walking mechanism is a two-wheeled differential chassis. There is an omni-directional wheel in the rear of the robot (see Fig. 3, each dark circle represents a sonar sensor). The velocity of the robot is set to be 0.1m/s.

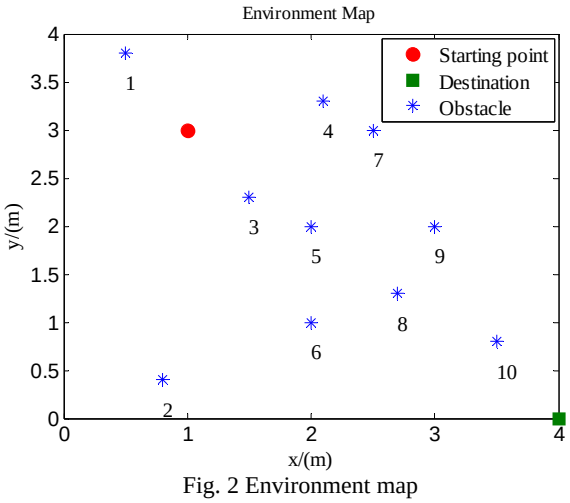

Fig. 2 Environment map

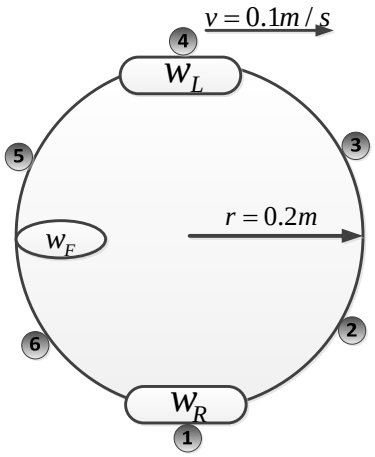

Fig.3 Top view of the robot structure

4.2 Experiment results

To compare the performance of both models, we choose the program running time which indicates the convergence speed of two models as the index. Obviously, it is the less the better. In every experiment, we repeat the program 100 times to eliminate the influence of accidental factors. No supervision is given in both experiments. The robots controlled by *OCLM* or *OCLMBP* completely follow Skinner's operant conditioning principles to navigate by means of sonar sensor information. During the navigation collisions are used to produce negative reinforcement signals while approaching destination is used to produce positive reinforcement.

Figures 4-7 show one example without collision for each robot (*OCLM* and *OCLMBP* robot respectively) in two experiments. Figure 8 shows the comparison of the running time. It can be seen that the *OCLMBP* robot has better performance in both experiment, no matter whether the positions of obstacles are fixed or not. *OCLMBP* as the revised version of *OCLM* is much faster in convergence than *OCLM*.

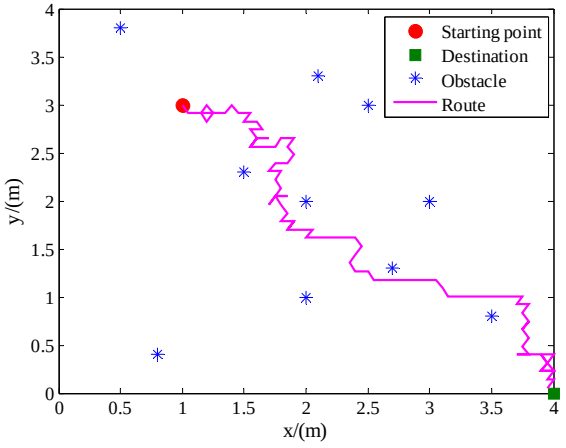

Fig. 4 One example of a route in the fixed position experiment performed by *OCLMBP* robot

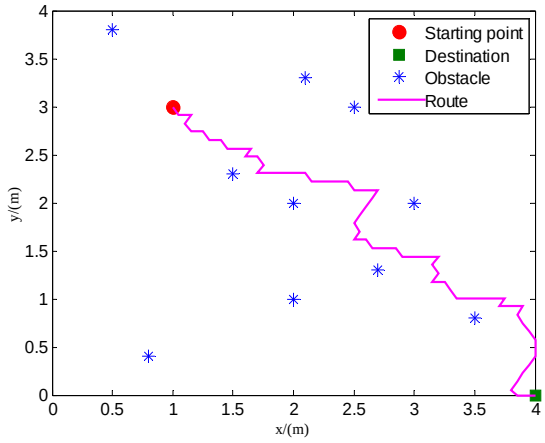

Fig.5 One example of a route in the fixed position experiment performed by *OCLM* robot

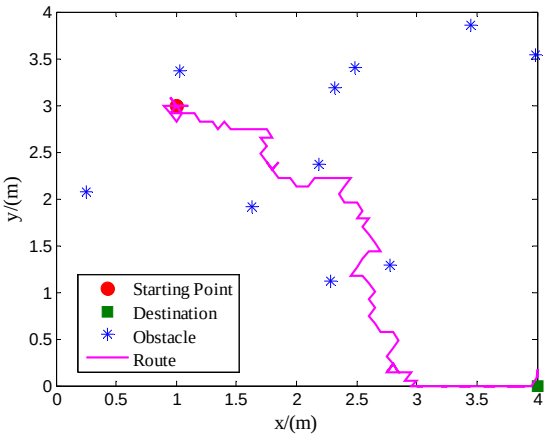

Fig.6 One example of a route in the random position experiment performed by *OCLMBP* robot

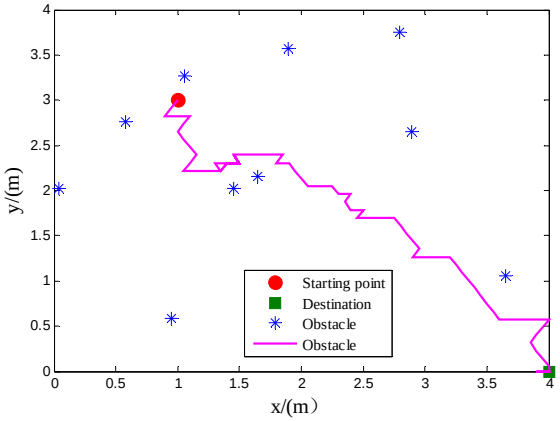

Fig.7 One example of a route in the random position experiment performed by *OCLM* robot

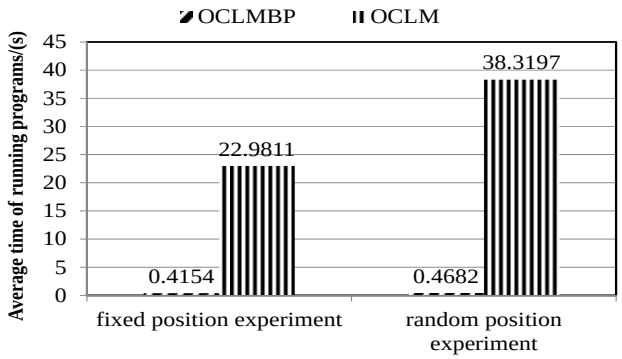

Fig.8 Average time of running programs in two experiments for both models

5 Conclusion

In this paper we have proposed a model which revises the original *OCLM* and realizes it in BP(back-propagation) network. The model named *OCLMBP* can learn obstacle avoidance for a wheeled mobile robot by means of sonar sensor information. The robot controlled by *OCLMBP* progressively learns to avoid the obstacles without any supervision, but by negative reinforcement signals produced by collisions and by positive reinforcement signals produced by approaching the destination, which totally obeys Skinner's operant conditioning rule and successfully reproduces animals self-learning behavior. To compare the performance of *OCLMBP* and *OCLM*, we design and implement the comparison experiments. The experiment results show that *OCLMBP* is much faster in convergence speed than *OCLM*, which indicates it gets better performance.

References

[1] R.Brooks, From earwigs to humans, *Robotics and autonomous Systems*, 20 (1997): 291-304,1997.

[2] J. Weng, et al, Autonomous mental development by robots and animals. *Science*, 291 (2001): 599-600, 2001.

[3] M. Asada, K. Hosoda, et al, Cognitive developmental robotics: A survey, *IEEE Transactions on Autonomous Mental Development*, 1 (1): 12-34, 2009.

[4] D. Stachowicz, G.M. Kruijff, Episodic-like memory for cognitive robots, *IEEE Transactions on Autonomous Mental Development*, 4(1):1-16,2012.

[5] A.G. Barto, S. Singh, and N. Chentanez, Intrinsically motivated learning of hierarchical collections of skills, in *Proceedings of the 3rd International Conference on Development and Learning*, 2004:112-119.

[6] B.F. Skinner, *The Behavior of organisms: an experimental analysis*, New York: D. Appleton-Century Company,1938.

[7] E. Zalama, P. Gaudiano, J.L. Coronado, *Obstacle avoidance by means of an operant conditioning model*, Berlin: Springer, 1995:471-477.

[8] P. Gaudiano, C. Chang, Adaptive obstacle avoidance with a neural network for operant conditioning: experiments with real robots, in *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 1997: 13-18.

[9] P. Gaudiano, E. Zalama, C. Chang, and J. Coronado, A model of operant conditioning for adaptive obstacle avoidance, in *From Animals to Animats*, Cambridge, MA,USA: MIT Press, 1996:373–381.

[10] H. Ishii, M. Nakasuji, M. Ogura, et al, Accelerating rat's learning speed using a robot-The robot autonomously shows rats its functions, in *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, Roman: IEEE Press,2004: 229-234.

[11] K. Itoh, H. Miwa, M. Matsumoto, et al, Behavior model of humanoid robots based on operant conditioning, in *Proceedings of 5th IEEE-RAS International Conference on Humanoid Robots*, Tsukuba: IEEE Press, 2005: 220-225.

[12] T. Taniguchi, T. Sawaragi, Incremental acquisition of behaviors and signs based on a reinforcement learning schemata model and a spike timing-dependent plasticity network, *Advanced Robotics*, 21(10): 1177-119, 2007.

[13] X.G. Ruan, J.X. Cai, L.Z. Dai,Compute model of operant conditioning based on probabilistic automata, *Journal of Beijing University of Technology*, 36(08): 1025-1030,2010.

[14] J.X. Cai, X.G. Ruan, OCPA bionic autonomous learning system and its application to robot poster balance control, *Pattern Recognition and Artificial Intelligence*,24 (01):138-146,2011.

[15] J. Huang et. al, Operant conditioning learning model in the bionic experiment, *Applied Mechanics and Materials*, 373(2013): 255-264,2013.

[16] J. Huang et.al, A Learning Model Based on Operant Conditioning Principles, *Control and Decision*, accepted.

[17] D. E Rumelhart, G. E. Hintont, and R. J. Williams, Learning representations by back-propagating errors, *Nature*, 323(6088), 533-536,1986.

[18] K. I. Funahashi, On the approximate realization of continuous mappings by neural networks, *Neural networks*, 2(3):183-192,1989.

[19] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems*, 2(4): 303-314,1989.