BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXI No. 3/2009	23 - 28	Seria Tehnică

Biologically Inspired Optimization Methods – A Review

Petia Koprinkova-Hristova

Institute of Control and System research, Bulgarian Academy of Sciences,
Akad. G. Bonchev str. Bl.2, Sofia – 1113, Bulgaria
e-mail: pkoprinkova@icsr.bas.bg

Abstract

The paper presents a short overview of the biologically inspired methods for optimization using neural networks methodology. The main focus is on dynamic programming optimization method and its forward approximations by adaptive critic design methods. Some recent parallels with the psychology and neuroscience achievements concerning reinforcement learning in the humans' brain are also considered. The possible future directions of investigations that could make a bridge between neuroscience and control theory are marked off.

Key words: *optimization, dynamic programming, neural networks, reinforcement learning*.

Introduction

The well known method of dynamic programming proposed by Bellman [2] offers common approach to finding the global maxima (minima) of a given quality function (criteria of optimality). However the number of calculations needed to solve given optimization task arises dramatically with its dimension. Moreover, the real world is nonlinear as a rule so the most models of the real objects under consideration are nonlinear that makes further more difficult the real application of the method.

One possible decision of the so called "curse of dimensionality" is offered by the "neuro-dynamic programming" [3] and "adaptive critic designs" [8]. The core of these approaches is usage of neural networks' ability to map quite complex nonlinear dependences and to learn from experience like living organisms.

The paper presents a short overview of the biologically inspired methods for optimization using neural networks methodology. The main focus is on dynamic programming optimization method and its forward approximations by adaptive critic design methods. Some recent parallels with the psychology and neuroscience achievements [4, 7] concerning reinforcement learning in the humans' brain are also considered. The possible future directions of investigations that could make a bridge between neuroscience and control theory are marked off.

Dynamic programming problem

Classical dynamic programming

The principal of the optimality given by Bellman [2] in 1957 says: "An optimal trajectory has the property to be optimal starting with any of its points till its end no matter how the chosen point is reached". Using this principal dynamic programming solves the optimization task starting from the end point backwards to its beginning in time. The overall method is described as follows [2, 3]:

The object under consideration is described as discrete dynamical system of equations:

$$x_{k+1} = f_k\left(x_k, a_k, d_k\right), \quad k = 0,1,\ldots,N-1 \tag{1}$$

where $x_k \in X$ is state variables vector, $a_k \in A$ is control variables vector and $d_k \in D$ is disturbances vector. For the given the initial state x_0 the purpose is to find such control policy $\left\{a_1, a_2, \ldots, a_{N-1}\right\}$ that minimizes (maximizes) the target function:

$$J_0\left(x_0\right) = \sum_{k=0}^{N} \gamma^k U_k\left(x_k, a_k, d_k\right) \tag{2}$$

where U_k is utility function or the "cost of the current step" and $0 \le \gamma \le 1$ is parameter.

The dynamic programming method solves the optimization task as follows:

$$J_N\left(x_N\right) = U_N\left(x_N\right) \tag{3}$$

$$J_k\left(x_k\right) = \min_{a_k \in A}\left\{U_k\left(x_k, a_k, d_k\right) + \gamma J_{k+1}\left(x_{k+1}\right)\right\} \tag{4}$$

The equality (4) is called Bellman's equation. It could be solved only by simulations. Hence with the increasing of the task dimension the so called "curse of dimensionality" aroused.

"Punish/reward" training and Q-learning

The reinforcement learning also called "learning without teacher" is introduced as method for artificial neural networks training. It arises as attempt to learn "by experience" rather than "by examples". The term "adaptive critic" or "training by critic" is introduced in [18] where it is demonstrated how a simple adaptive linear element can "learn" a strategy starting without a-priori information. Further in [1] two neuron-like elements are proposed that are able to learn to balance inverted pendulum using "punish/reward" signal only. The method mimics the living organisms' behavior called "action – error". It is based on trying actions and receiving environmental reaction about how good or bad was their decision. The main learning principal that arises from animal's conditioning is: "if the current action is followed by a positive reaction from the environment that the tendency of repeating that action in future is strengthen or rewarded; in other case this tendency is decreased or punished". This approach has two versions – associative and non-associative. While in the non-associative version the "actor" (an element or controller that takes decision and undertakes actions) receives from the environment only simple signal indicating his action as "good" or "bad", in the associative version there is additional information about the current environmental conditions as well. A further version of that "punish/reward" learning is called Q-learning [13]. It consists of prediction of discounted sum of the future rewards r_k:

$$Q_k(x_k, a_k) = \sum_{j=0}^{\infty} \gamma^j r_{k+j} \tag{5}$$

in order to try to maximize it. In that case the optimization task is presented as:

$$Q(x_k, a_k) = r_k + \gamma \max_b Q(x_{k+1}, b) \tag{6}$$

where $0 \le \gamma \le 1$ is parameter. As it is easily seen the equation (6) is analog to the Bellman's equation (4).

This method however needs big amount of memory to keep all the possible situations and their possible rewards or punishments.

Temporal difference and error backpropagation

The next step towards predication of future rewards is the method of "temporal differences" [11]. The main idea is to try to mimic animals' ability to predict future outcomes on the basis of their previous experience without waiting the final results of their actions in the future. It is related to Hebbian learning in cortical hippocampal synapses [5] thus making the bridge between psychology, computer science and engineering. The main idea is in defining of the adaptive critic's output error as:

$$E(k) = \gamma J_{\kappa+1} + U_\kappa - J_\kappa \tag{7}$$

This means that at each time step the critic will predict future values of the optimality criteria having information of its current value and the current utility value. It allows solving the optimization task in forward manner.

One of the most powerful and widely used training algorithms for neural networks is error backpropagation method [9, 14]. Although developed especially for neural networks, the method is quite common and can be applied to any system of calculations that could be ordered a "chain". The method gives a rule for calculation of a given function's derivatives with respect to all the variables included in such chain no matter whether these variables are included explicitly in or not in that function. So the error backpropagation and especially its recurrent version [16] can be applied to optimization tasks such as dynamic programming.

Adaptive critic designs as approximation of dynamic programming

The overall approach called "neuro-dynamic programming" comprises the neural networks for approximation of Bellman's equation right part or/and needed models of the actor, environment etc. and behavioral approaches for predictor's (called also adaptive critic) training.

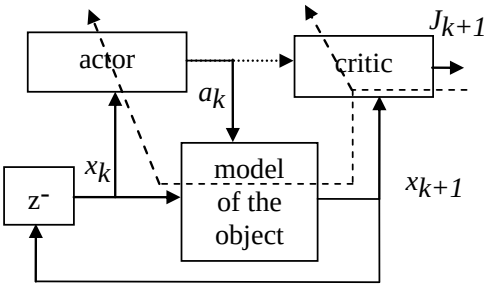

Fig. 1. Adaptive critic design with known object's model.

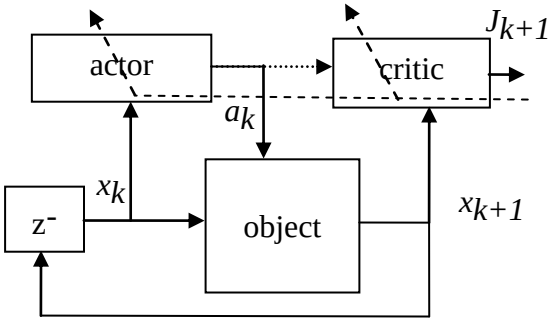

Fig. 2. Adaptive critic design without known object's model.

The adaptive critics are classified as follows [8, 15, 17]:

a) according to their inputs:
   - action independent –at the critic's input is presented only information about the environment's current state;
   - action dependent – the information about actions taken is also in presented at the critic's input.

b) according to their output variables:
   - heuristic dynamic programming – the output of the critic is performance criteria value at current moment;
   - dual heuristic programming – the output of the critic is the first derivative of the performance criteria;
   - global dual heuristic dynamic programming – both performance and its derivative are critics output.

The training schemes of adaptive critic designs depend on the presence or absence of a model of controlled environment [6, 10] – see figures 1 and 2. In both cases the error backpropagation method is appropriate. In the case of critic training temporal difference error is taken while in the case of actor (controller) training, the performance function value is minimized (maximized).

The next step - Hierarchical reinforcement learning

Although the adaptive critic design method offered a forward approach to dynamic programming optimization, it still suffers from scaling problem – the bigger is the number of environment's states to be explored by the actor and the number of its allowed actions the more time will need the critic to be trained adequately. So in [12] the hierarchical reinforcement learning was proposed. The core of that approach is introduction of the so called "abstract actions" that unite several consecutive simple actions into a policy to be followed, several sub-goals that have to be reached before final goal (optimum of the overall performance criteria) and related to them pseudo-rewards. In this way the training time can be dramatically decreased. However the choice of such abstract actions is tricky and depends on specificity of the optimization task to be solved. The sub-goals are usually at the "bottlenecks" of the states' space and have to be chosen also task-dependent.

Neuroscience parallels

Recently in the literature appeared new works oriented to psychology [4, 7] that consider the backward relations from the reinforcement learning to the human's decision-making analysis. They related the Pavlovian conditioning to the prediction learning, i.e. critic's work and the instrumental conditioning to the learning how to select actions that will increase future outcomes (rewards), i.e. actor's work. Further more they've made several parallels between hierarchical reinforcement learning and hierarchical structure of humans' behavior.

In that way the authors have made "a bridge" between neuroscience and optimal control theory. Further more they suggested to explore the human's behavior using the reinforcement learning normative model and to consider discrepancies between real neuroscience data obtained by contemporary methods (such as functional magnetic resonance imaging, electrophysiological recordings and neuro-modulators' functions) and reinforcement learning model as a good basis for new knowledge achievements.

The comparison of experimental and computational data stated interesting new directions of research both from psychological and computational points of view. Some of them are:

- How does learning from one task affect subsequent learning of another one?
- Which are the structures in the brain that are equivalent to the elements of the reinforcement learning normative models?
- How could be explained some of the contemporary data of brain's investigation with the purely theoretical reinforcement learning paradigms?

Conclusions and directions for future work

The discussed here biologically inspired optimization methods are a part of much more wide area of research both from control theory and from psychological point of view. The present review doesn't pretend to be exhaustive rather than to point out some of the milestones in this area and the interesting relations between these completely different areas of science that appeared to merge in the near future.

From my point of view it will be interesting to try to use unexplained neuro-physiological phenomena related to human's decision-making to develop new optimization techniques accounting for these data in a better way. This may lead to new achievements in control theory allowing us to make one step ahead towards practical artificial intelligence application.

Acknowledgments

This work is partially supported by the Bulgarian Science Fund under the project No TN 1509/05 "Control of Mixed Culture Fermentations in Biochemical and Food Industries" and bilateral project "Monitoring of biotechnological and ecological processes for quality control in the food industry" between ICSR - BAS and "Lucian Blaga" University of Sibiu

References

1. Barto, A.G., Sutton, R.S., Anderson, C.W. – *Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems*. IEEE Trans. on Systems, Man and Cybernetics, vol. 13, No 5, 1983, pp.834-846.
2. Bellman. R.E. – *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

3. Bertsekas, D.P., Tsitsiklis, J.N. – *Neuro-Dymanic Programming*. Athena Scientific, Belmont, MA, 1996.
4. Botvinick, M.M., Niv, Y., Barto, A.G. – *Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective*. Cognition, 2008, doi:10.1016/j.cognition.2008.08.011 (in press).
5. Dayan, P. – *Matters temporal*. Trends in Cognitive Sciences, vol.6, No3, March 2002, pp.105-106.
6. Horvath, G. – *Neural Networks in System Identification*. In: Neural Networks for Instrumentation, Measurement and Related Industrial Applications, Edited by S. Ablameyko, L. Goras, M. Gori and V. Piuri. NATO Science Series, vol.185, IOS Press, Amsterdam, 2003, pp.43-78.
7. Niv, Y. – *Reinforcement learning in the brain*. Journal of Mathematical Psychology, 2009, doi:10.1016?j.jmp.2008.12.005 (in press).
8. Prokhorov, D.V. – *Adaptive critic designs and their applications*. Ph.D. dissertation, Department of Electrical Engineering, Texas Tech. Univ., 1997.
9. Rumelhart, D.E., McClelland, J.L. – *Parallel Distributed Processing*. Vol. 1, MIT Press, Cambridge, MA, 1986.
10. Si, J., Wang, Y.-T. – *On-line learning control by association and reinforcement*. IEEE Trans. on Neural Networks, vol.12, No2, 2001, pp.264-276.
11. Sutton, R.S. – *Learning to predict by methods of temporal differences*. Machine Learning, vol.3, 1988, pp.9-44.
12. Sutton, R.S., Precup, D., Singh, S. – *Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning*. Artificial Intelligence, vol.112, pp.181-211, 1999.
13. Watkins, C., Dayan, P. – *Q-learning*. Machine Learning, vol.8, 1992, pp.279-292.
14. Werbos, P.J. – *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Ph.D. dissertation, Committee on Applied Mathematics, Harvard Univ., Cambridge, MA, 1974.
15. Werbos, P.J. – *Building and Understanding Adaptive Systems: A Statistical/Numerical Approach to Factory Automation and Brain Research*. IEEE Trans. on SMC, vol. 17, No 1, 1987, pp.7-20.
16. Werbos, P.J. – *Backpropagation Through Time: What It Does and How to Do It*. Proceedings of the IEEE, vol. 78, No 10, 1990, pp.1550-1560.
17. Werbos, P.J. – *Neurocontrol and Supervised Learning: An Overview and Evolution*. In: Handbook of Intelligent Control, Ed. D. White and D. Sofge, Van Nostrand, 1992.
18. Widrow, B., et al. – *Punish/Reward: Learning with a Critic in Adaptive Threshold Systems*. IEEE Trans. on SMC, vol. 3, No 5, 1973, pp.455-465.

O prezentare generală a metodelor de optimizare inspirate din biologie

Rezumat

În lucrare se realizează o scurtă prezentare a metodelor de optimizare inspirate din biologie şi care utilizează metode bazate pe reţele neuronale. Principalul obiectiv se concentrează pe metodele de optimizare dinamică şi aproximările iniţiale asociate, determinate prin metode de proiectare adaptiv-critice. Sunt de asemenea considerate unele abordări recente care privesc paralelismul, cu realizări din psihologie şi neurologie referitoare la procesele care au loc în creierul uman la învăţarea întărită. Investigaţiile viitoare se vor concentra pe identificarea unor posibile punţi de legătură între procesele neuropsihologice şi teoria controlului optimal.