IMPLEMENTING VARIABLE STRATEGY IN REAL-TIME REFLEX DECISION-MAKING AND CONTROL

Baofu Duan and Yoh-Han Pao

Case Western Reserve University
Cleveland, Ohio 44106, U.S.A

Abstract: In principle, autonomous decision-making and control can be implemented efficiently in reflex mode with use of neural networks. In practice there are situations where the very strategy of decision or control needs to be varied in accordance with circumstances. For example, in multi-objective control, emphasis may need to be shifted from one objective to another in accordance with system state. Similar considerations apply to decision-making. This paper reports on how knowledge of such variation in strategy can be captured and implemented. Control of the inverted-pendulum/cart task is used to illustrate the approach. Other applications are also discussed. *Copyright 2000 IFAC.*

Keywords: Adaptive Control, Autonomous ,Control, Self-adaptive Control, Neural Networks.

1.INTRODUCTION

The results of investigations in many disparate fields of research indicate that intelligent behavior is associated with hierarchical structure.

In such systems, reaction times and extent of abstraction increase with increase in level in the hierarchy. In addition, the lower levels might be characterized by continuous-variable dynamics and the higher levels by logic-driven decision-making discrete-valued mechanisms. The interaction of these different levels, with their types of information, leads to a 'hybrid' system (Branicky, 1995).

The conventional wisdom is that the lower levels of such systems are likely to be populated by fast, capable, stimulus-response reflex elements. This present paper is concerned with an important but subtle aspect of the behavior of reflex actuators or decision-makers, that of adaptation in the strategy of response.

This is a fine point but an important one. It is not that the reflex is only capable of one fixed motion or class of decisions. In fact, all useful reflexes are generally infinitely variable and appropriate in response, acting in accordance with some learned pattern acquired from experience or prescribed by rules. The issue is that for some tasks, perhaps for all but the simplest 'knee-jerk' responses, the 'algorithm' itself at so-called higher levels also needs to be varied in response to changes in the task environment.

Specifically, in multi-objective control, the relative emphasis placed on the various individual objectives sometimes need to be changed in accordance with the system state and with changes in task environment. In a board game such as Othello, for example, the strategy guiding the choice of move might fluctuate between 'maximum disc' and 'least mobility' strategies, of which more will be said in the following.

The point is that there can be a shift of emphasis or a change in the identification of operative governing rules as a task proceeds. We feel that it is important to understand how to avoid or delay having to resort, prematurely, to the higher logic-driven entities in a hierarchical system. In fact the results of this investigation suggest that some rules might very well better be implemented in connectionist mode as reflexes.

To fix ideas and to illustrate some points, we describe the task of balancing and positioning an inverted pendulum/cart system. The pendulum is hinged and is constrained to swing about that hinge in a one-dimensional angular motion in a plane. The hinge is mounted on a cart and the cart is free to move on a platform. There is evidence to suggest that the inverted pendulum can be maintained upright through application of impulses at appropriate times at one end or other of the cart. However if one is not

extremely skilled, the cart will eventually roll off the platform which is of finite length, or the pendulum will collapse onto the cart. This is especially so if there are no frictional forces, neither in the hinge nor in the cart motion.

This is a well-known task and different control strategies have been used in the various previous studies (Bryson and Luenberger, 1970, Mori et al., 1976, Barto et al., 1983, Widrow, 1987, Nyberg and Pao, 1995).

Qualitatively speaking, the pendulum moves very slowly when it is nearly upright and the cart changes momentum instantly upon application of impulse. In addition, there are regions in system state space where application of impulse reduces both angular and positional deviations and there are other regions where improvement in attainment of one objective is accompanied by deterioration in the other. Given those insights, it is possible to devise control systems that will carry out the combined task for an indefinitely long time, for certain regions of the 4-dimensional state space.

Our investigation shows that reflex control with a fixed objective function is limited and brittle. But with incorporation of adaptive shift of emphasis between the two parts of the overall objective function, reflex control is strengthened, and regions of prior difficulty can be brought into control. The issue is how to acquire a representation of the manner in which strategy is to be changed and how that knowledge might be generated and used in control.

2.THE INVERTED PENDULUM/CART TASK

The inverted pendulum/cart task has been used as a useful laboratory idealization of inherently unstable systems that can be maintained in dynamically stable condition through use of appropriate control action. In this task, a controller is asked to learn how to balance an inverted pendulum and at the same time control the position of the cart on which the pendulum is mounted. Different versions of the inverted pendulum exist, but the most common is a pole hinged to a moving cart as shown in Figure 1. The pendulum is hinged so that it is constrained to swing about the hinge in a one-dimensional angular motion in a plane. The hinge is mounted on a cart and the cart is free to move on a platform in a straight line.

The inverted pendulum/cart system can be modeled by the two nonlinear differential equations (Cannon, 1967):

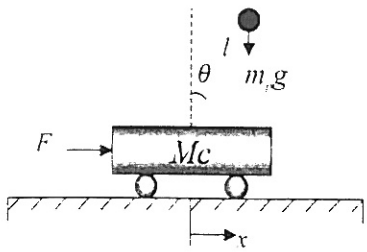

Fig. 1. The Inverted-Pendulum Cart Task

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta \dfrac{-F - m_p l\dot{\theta}^2 \sin\theta + \mu_c \operatorname{sgn}\dot{x}}{M_c + m_p} - \dfrac{\mu_p \dot{\theta}}{m_p l}}{l\left(\dfrac{4}{3} - \dfrac{m_p \cos^2\theta}{M_c + m_p}\right)} \quad (1)$$

$$\ddot{x} = \frac{F + m_p l(\dot{\theta}^2 \sin\theta - \ddot{\theta}\cos\theta) - \mu_c \operatorname{sgn}\dot{x}}{M_c + m_p} \quad (2)$$

where θ is the angle between the pendulum and vertical, x is the horizontal position of the cart, g = $9.81 m/S^2$ is the acceleration due to the gravity, M_c = $1.0kg$ is the mass of cart, m_p = $0.1kg$ is the mass of pendulum, l = $1.0m$ is the half length of pendulum, F is the force applied to cart's center of mass, μ_c is the cart friction coefficient, and μ_p is the hinge friction coefficient. In our present approach, μ_p and μ_c are set to be 0, i.e., no friction is involved. In other words, the cart moves freely on the platform without any restraining force and the pendulum swings freely about its hinge without any restraining torque. This makes it more difficult to attain the multiobjective balancing and positioning act.

These equations are assumed to be not known and they were only used to simulate the dynamics of the system. Euler's method with a sampling time interval h = $0.05Sec$ (second) is used for simulation. The sampling rate of the system's state and the rate at which control forces are applied are the same as the basic simulation rate, i.e., $20Hz$.

For convenience, an impulse $I = F \cdot h$, is used instead of force to denote the action on the cart. The impulse has a value between -0.5 *Newton·Sec.* and 0.5 *Newton·Sec.* since the value of the force is limited in an interval [-10*Newton*, 10*Newton*]. The total length of the platform is $2m$ and the range of x is from $-1m$ to $1m$. Similar to most of the previous approaches, the pendulum angle values of interest are limited to be within a range, chosen to be [-10°, 10°] in our implementation.

3. FIXED STRATEGY CONTROL

In fixed strategy mode, a simple neural-net controller is implemented to control the inverted pendulum/cart

task. The configuration of the system is described in Fig. 2.

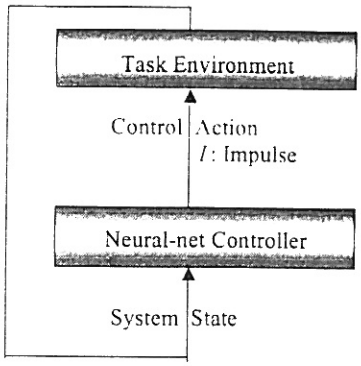

Fig. 2. A simple neural-net controller

The training of the controller is through an evolutionary process. At the start, the parameters of the controller are generated randomly. Then for each randomly generated system state, several different values of impulses are tried to decide what action should be generated. Then the difference of the actual output of the controller and the desired value of impulse is used to update the parameters of the controller.

The objective function used for training is defined as the square of the difference between the actual output of the environment and desired system state. In this implementation, the objective function is defined as:

$$E(\theta,x) = \alpha\theta^2 + (1-\alpha)x^2 \qquad (3)$$

where α is a handcrafted constant that is used to adjust for the scalar differences between θ and x.

In the objective function, θ and x are used to denote the angular and positional deviations since the desired values of the pendulum angle and the cart position are both 0.

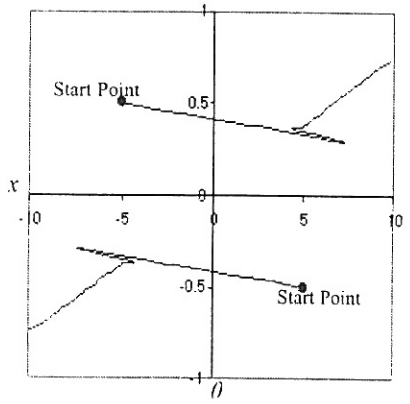

Fig. 3. Two controlled trajectories using a simple neural-net controller

Two controlled trajectories using a simple neural-net controller is shown in Fig. 3, where the controller is learned with use of $\alpha = 0.80$ in the objective function. For one trajectory, $\theta = -5°$ and $x = 0.5m$ at the start point, whereas $\theta = 5°$ and $x = -0.5m$ at the start point for the other one. For both controlled trajectories, it is assumed the system starts with zero angular velocity and zero position velocity. The simulation results show that the pendulum will swing out of the control range, regardless of which initial stationary position is used as the start point.

In our investigation different controllers were trained using different values of the parameter α in equation (3). However, none of them could maintain the pendulum in a nearly upright position and also keep the cart on the platform, even though the controllers indeed acted optimally all the time. However, it is interesting to see that the system trajectories change with the parameter α. This implies a possibility of modulating reflexes by changing the value of the parameter α in accordance with changes in system states. In the next section, we will introduce a better method of control based on using variable values of the parameter α in the objective function. Simulation results show that the system can be controlled for an indefinitely long period of time if appropriate reflex control modes are used adaptively.

4. ADAPTIVE STRATEGY CONTROL

If this task were to be handled in hybrid system manner with discrete-valued decision-making at the higher level there might be need for two modes of operation, in addition to others, for dealing with extreme situations. In one extreme there would be little consideration of the position of the cart, all attention would be devoted to keeping the pendulum at a reasonably small angle from the vertical. In the other extreme, all attention would be centered on bringing the cart reasonably close to the center of the platform and maintaining it in that condition. In practice, what is required is a strategy and an ability to vary the emphasis depending on the circumstances, so as to maintain the pendulum in upright position or at a feasible angle and with the cart on the platform for as long a period of time as possible.

In this paper we describe a methodology for acquiring and implementing such a variable control strategy. That is, we acquire knowledge of what the correct strategic aspects of control should be for several instances and then build an interpolating model for generating correct control actions for all circumstances.

In equation (3), the parameter α is used to adjust for scaling differences between pendulum angle and cart position. On the other hand, the parameter α can also be viewed a modulation parameter used to shift the controller's focus of attention. That is it could be

used to indicate how much emphasis should be put on the pendulum angle and how much on the cart position. In one extreme case, α equals to 1 and there would be no consideration of the position of the cart, all attention would be devoted to keeping the pendulum at the desired angle. In the other extreme, α equals to 0 and all attention would be given to bringing the cart reasonably close to the center of the platform and maintaining it in that condition. The value of the parameter α is usually between 0 to 1. However, if we choose α larger than 1 or less than 0, there is a different meaning for this parameter. In the objective function, if α is larger than 1, the position deviation part will be negative. So the system will try to increase x instead of decreasing it as usual. If α is less than 0, the angle deviation part in the objective function will be negative. So the system will try to increase θ instead of decreasing it as usual. The fact that a system is trying to move away from its objective can be explained as that the system is taking risks on θ or x.

This suggests that a variable objective function, with α value depending on the system state, can be used. The objective function will be written in following as:

$$E(\theta,x) = \alpha(s)\theta^2 + (1-\alpha(s))x^2 \qquad (4)$$

where s is the system state.

It is a novel idea that a variable objective function can be used in the control system. The objective function is dependent on the system state so that a controller can shift its focus of attention in different scenarios. Another important idea is that the sometimes risks are taken not to make the performance thrilling, but as interim measures to move the system to more favorable states for ultimate attainment of convergence.

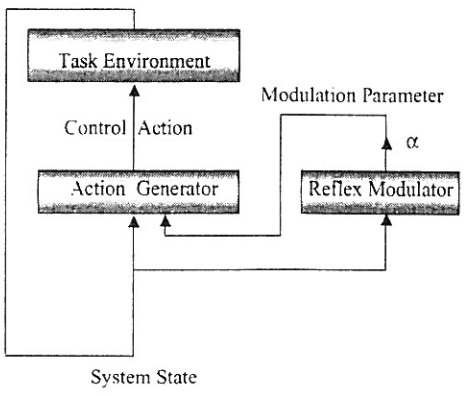

Fig. 4. Structure of an Adaptive Strategy Controller

Based on these ideas, the proposed adaptive strategy controller is designed as in Fig. 4. It includes two parts: an action generator and a reflex modulator. The action generator generates control action, the value of the impulse and its direction, when presented with the system state vector and a modulation parameter as input. The reflex modulator generates the value of the modulator parameter, which is used as an input of the action generator so as to modulate the behavior of the action generation.

The action generator is learned through evolution. At the start, the parameters of the action generator network are generated randomly. Then for each randomly generated state, several different modulation parameters are generated. For each modulation parameter α, an optimal value of the control action is found as the desired value. Then the action generator can update its parameters by comparing its actual output with the desired value of impulse.

The reflex modulator may also be learned through an evolutionary process. A reflex modulator network would be generated randomly at first or in accordance with a few examples of inputs and associated outputs. Then the reflex modulator is incorporated into the environment with the action generator and its long-term performance is evaluated. Then a set of additional reflex modulators are generated by randomly changing some parameters of the current reflex modulator. The long-term performance of these reflex modulators are also evaluated. The reflex modulator with best performance is chosen to generate new reflex modulators.

One successful controlled trajectory of pendulum/cart using the optimized adaptive strategy controller is shown in Fig. 5. At the start point $\theta = 5^\circ$, $x = 0.5m$, angular and position velocities are assumed to be 0. It is shown that with use of the proposed control configuration, the system does converge efficiently to the objective.

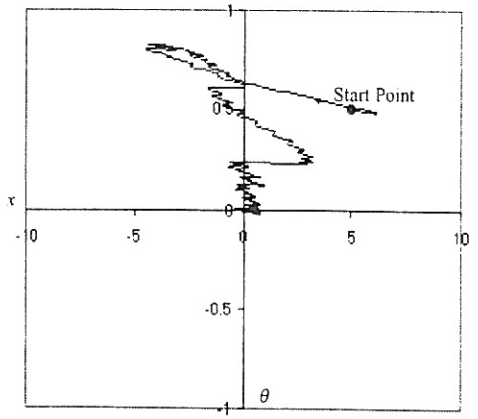

Fig. 5. One Successfully Controlled Trajectory

5. OTHER DOMAINS OF APPLICATION

It would seem that the general theme of adaptive modulated reflex control, trained through experience, may be valid and useful for a number of applications including decision-making, game-playing, image interpretation and so on. To fix ideas we describe how that approach might be used to help an autonomous software agent develop a winning strategy for a game.

The game of Othello is a delightful board game. The rules are simple and even beginners can enjoy the game but the game is actually quite complex and can be used to explore paradigms for learning strategy and so on. It is played on a 8×8 board with a set of dual-colored discs. Starting with 4 discs in compact diagonal array in the center of the board, one diagonal being black and the other white, the two players take alternate turns to play a disc each turn. The objective is to bracket some of the opponent's pieces in some array, vertical or horizontal, or along one or more diagonals. Bracketed discs are lost to the opponent and are turned over to show the player's color. A move is legal only if some bracketing is achieved by that move, otherwise the player is said to have no mobility, no ability to move, and misses a turn. The game ends with all 60 pieces have been played on the board or if neither side is able to move. Each side is allowed 30 minutes in which to make all of its 30 moves (Rosenbloom, 1982, Billman and Shaman, 1990, Moriarty and Miikkulainen, 1995). Highly competent computer programs have been evolved for playing the game of Othello (Rosenbloom, 1982, Billman and Shaman, 1990, Moriarty and Miikkulainen, 1995). There are at least two 'pure' strategies one being the 'maximum disc' and the other the 'least mobility'. In the former one plays to maximize the number of opponent's discs captured and flipped. With the other strategy one plays so as to give the opponent as few legal moves as possible on his next turn.

It seems that even the pure form of 'least mobility' is a formidable strategy. But it is not all and it may be implemented in different ways. Perhaps the 'maximum disc' is appropriate at early stages of the game, except for certain circumstances, but increasing importance might be given to 'least mobility' as the game proceeds.

This is cited as an example of what we mean when we advocate exploring the design of an adaptive modulator for reflex control. In this case, the modulator would take a look at the state of the board and issue advice on strategy, whether it should be 'maximum disc' or 'least mobility' or a mixture of the two, and how the 'mixing' might be done. (The interested reader might also visit some web sites such as http://www.maths.nott.ac.uk/othello/othello.html).

In other instances human judgement is effective in effecting a high quality of task performance but it is also very difficult to capture that human judgement. The adaptive modulation method might be used to capture a representative of such judgment and strategies.

REFERENCES

Barto, A.G., Sutton, R.S. and Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-13**, 834-846.

Billman, D. and Shaman, D. (1990). Strategy knowledge and strategy change in skilled performance: a study of the game Othello. *American Journal of Psychology*, **103**, 145-166.

Branicky, M.S. (1995). *Studies in Hybrid Systems: Modeling, analysis and control*. MIT D.SC. thesis.

Bryson, A.E. and Leuenberger, D.G. (1970). The synthesis of regulator logic using state-variable concepts. *Proceedings of the IEEE*, **58**, 1803-1811.

Cannon, R. H. (1967). *Dynamics of Physical Systems*, McGraw-Hill, NY.

Mori, S., Nishihara, H. and Futura, K. (1976). Control of unstable mechanical system, control of pendulum. *International Journal of Control*, **23**, 673-692.

Moriarty, D.E. and Miikkulainen, R. (1995). Discovering complex Othello strategies through evolutionary neural networks. *Connection Science*, **7**, 195-209.

Nyberg, M. and Pao, Y.H. (1995). Automatic optimal design of fuzzy systems based on universal approximation and evolutionary programming. In: *Fuzzy Logic and Intelligent Systems* (Li, H. and M. Gupta (Ed)), 342-358. Kluwer Academic Publishers, Boston.

Rosenbloom, P.S. (1982). A world-championship-level Othello program. *Artificial Intelligence*, **19**, 279-320.

Widrow, B. (1987). The original adaptive net Broom-balancer. *IEEE International Symposium on Circuits and Systems*, **2**, 351-357.