In: Neurocomputing Research Development
Editor: Hugo A. Svensson, pp. 181-200
ISBN: 978-1-60021-931-3

Chapter 7

BASAL GANGLIONIC LEARNING APPLIED TO CONTROL OF AN UNDERACTUATED SYSTEM

Sungho Jo[1*] and Zhi-Hong Mao[2**]
[1]Media Laboratory, Massachusetts Institute of Technology, 20 Ames Street E15-054, Cambridge MA 02139
[2]Department of Electrical and Computer Engineering, University of Pittsburgh, 434 Benedum Hall, Pittsburgh PA 15261

Abstract

This chapter presents a learning algorithm inspired by the neural computation in the basal ganglia and cerebellum. The learning algorithm can be described as a switching-based reinforce algorithm. The performance of the learning algorithm is evaluated by an application to an underactuated system, the Cart-Pole system. Two separate controls, swing-up and balancing controls, are required to successfully achieve the task goal. Adaptive tuning of the control gain is performed by the learning algorithm. The control signal processing and reward-based learning rule can be interpreted in terms of basal ganglionic neural computations: The parallel operation of the two separate controls mimics the parallel signal processing of the direct and indirect pathways in the basal ganglia. The reward-based update rule may correspond to the long-term potentiation and long-term depression of synaptic plasticity in the striatum inducted by dopamine release. Switching between the two controls can be explained with a plausible neural operation in the cerebellum: Cerebellar neural circuit with inhibition can be interpreted as competitive signal processing, which attains a switching mechanism.

1. Introduction

This chapter studies a learning algorithm based on the stochastic reinforcement learning. Prior to the proposal of stochastic reinforcement learning, Hebbian learning had been a worldwide famous rule of learning that reveals an adaptation mechanism of the presynaptic and postsynaptic neurons. However, this synaptic update rule cannot guarantee the computational

* E-mail address: shjo@media.mit.edu
** E-mail address: maozh@engr.pitt.edu

performance of the whole network of neurons [Bartlett and Baxter 2000]. In contrast, the learning rule of backpropagation has gained popularity for computational efficiency in real applications of artificial neural networks, but this rule is not biologically plausible. Recently, a new synaptic-update rule, the stochastic reinforcement learning, has been investigated [Seung 2003; Bartlett and Baxter 2000; Bern and Sejnowski 1994; Florentin and Porr 2005]. The scheme of stochastic reinforcement learning seeks maximizing the long-term average of a reward signal by a collection of spiking neurons as parallel multi-agents. This update rule is computationally effective, and at the same time it is biologically plausible because it mimics some aspect of reinforcement learning in the basal ganglia.

This chapter tries to interpret the learning algorithm from the perspective of neural computation. The algorithm will be applied to the Cart-Pole problem, a problem of controlling the inverted pendulum on a cart. This problem is an old and challenging problem to demonstrate the effectiveness of control systems in analogy with the control of many real systems. Therefore, the Cart-Pole problem is popularly invited to evaluate a computational learning scheme.

Before the introduction of the main algorithm in the next section, some background material is reviewed in the following subsections.

1.1. Basal Ganglia

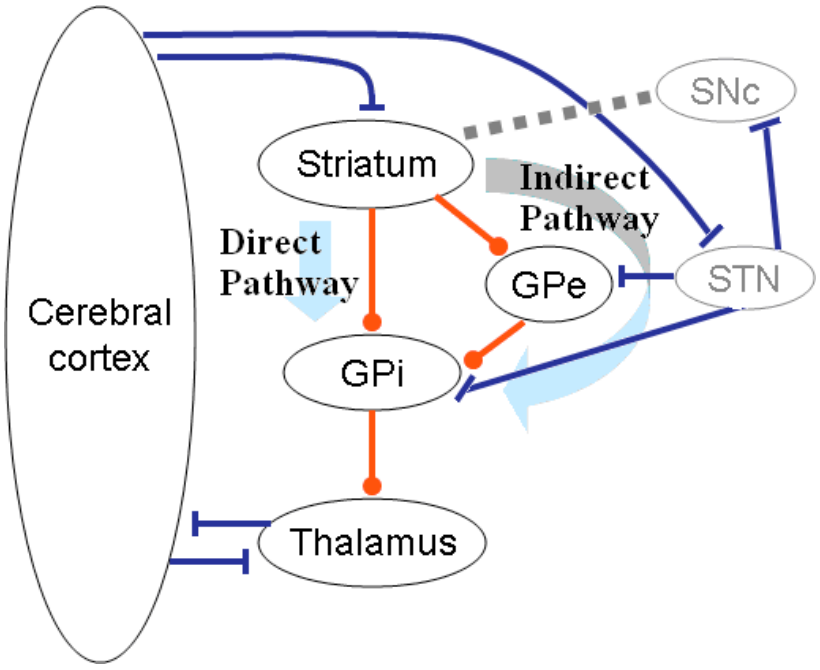

Figure 1. Direct and indirect pathways of the basal ganglia (This figure simplifies the connections).

The basal ganglia (BG) consist of four major nuclei: the striatum, the globus pallidus (GP), the substantia nigra(SN), and the subthalamic nucleus(STN). The striatum is the major recipient of inputs to the basal ganglia. Almost all areas of the cerebral cortex send excitatory

projection to the striatum. The major type of projection neuron is the medium size spiny neuron. The spiny neurons project inhibitorily to the GP and SN. The GP is divided into internal and external segments, i.e., globus pallidus pars interna (GPi) and globus pallidus pars externa (GPe). The GPi is the inhibitory output nuclei of the basal ganglia. The important neuroanatomy is that two parallel pathways modulate the output of basal ganglia: direct and indirect pathway as shown in Figure 1. It is widely accepted that the two pathways have opposing effects on the basal ganglia output nucleus. Activation of the direct pathway suppresses the neurons in the GPi, thus disinhibits the thalamus and increases thalamocortical activities. In contrast, the indirect pathway increases inhibition of the thalamus, thereby decreasing thalamocortical activities.

The SN consists of the pars reticulate (SNr) and pars compacta (SNc). The SNr is another BG output nucleus. The SNr is composed of large neurons that receive similar patterns of inputs as those of the GPi. The primary difference between the SNr and GPi outputs is that the lateral portion of SNr is connected with the areas of the cortex and of the brain stem that are involved in eye movement control [Mink 1996] (SNr is not shown in Figure 1). The SNc is a densely celluarly region contraining dopamine cells, which receives GABAergic and inhibitory input from the striatum. The SNc dopamine neurons also project back to the striatum. The dopamine pathway as well as the reciprocal connection between the striatum and SNc are thought to play a crucial role in learning carried by the BG. The STN receives an inhibitory, GABAergic input from the GPe, and an excitatory, glutaminergic input from motor cortex. The output from the STN is excitatory and glutaminergic and projects to GPi, GPe, and SNr.

The BG circuit can be considered as a big loop starting with inputs from multiple cortical areas to the BG and then returning, via thalamus, to the cortical areas [Nolte 1999]. This big loop can be further divided into multiple parallel loops. All these loops are similar in principle, but each uses different cortical areas and a distinctive portion of the striatum and globus pallidus. The cortico-BG-thalamocortical loop that corresponds to the putamen, is termed motor loop which basically participates in the control of movements. More recently, the motor loop has been considered to be further partitioned into oculomotor and several skeletomotor circuits. Each of these circuit divisions appears to derive input preferentially from the same areas of cerebral cortex to which they project. Thus, the BG loops appear to operate largely as parallel loops with little intercommunications [Kelly and Strick 2004]. The sub-architecture of each loop is very similar. It is widely felt that the BG provides the same or similar computational processing for different regions of the cerebral cortex.

1.2. Reinforce Algorithm

The REINFORCE algorithm is the online learning of an input-output mapping through a sequence of trial and error so as to maximize a statistical performance criterion. The algorithm is principally based on stochastic gradient ascent in a policy space, and does not require an explicit model of interacting environment or an explicit value function. As Williams indicated [Williams 1992], the REINFORCE algorithm is a class of reinforcement learning connectionist networks. Connectionist networks indicate parallel distributed signal processing. Each process unit receives external inputs from environment, and propagates outputs to environment after the work of activation function.

The unit is comparable with a neuron, and is called an agent in the world of reinforcement learning. The activation function describing the activity of the cell is an object to be designed. In the reinforcement learning, a policy to choose an action is equivalent to the activation function, where its output is the action and its input is the observed state information. An individual neuron communicates with many neurons in neighborhood. The effect of the connectionist networks can be represented by a weight vector. The weight vector corresponds to synaptic connections between neurons. Adjusting the weight vector represents synaptic plasticity, which is comparable with a learning algorithm. To assign a policy over actions to observed information, many possible policy strategies could be applicable. A stochastic policy is among the potential policy schemes. The stochastic policy is a probability of selecting an action based on observation. When the action value is continuous, the stochastic policy denotes a probability density function. The stochastic policy makes it possible advantageously to design a learning algorithm with no explicit information about environment and no approximation of a value function, and is applicable even with partially observable information (e.g., Partially Observable Markov Decision Processes).

Let X_t and a_t denote the observed information and action at time t respectively, and $\pi(a_t, X_t, W)$ represent the probability of taking an action a_t based on observed information X_t. The vector W consists of the weights representing the strength of connections in the neural networks. Because the policy π is comparable with activation function, each element of W is comparable with synaptic connection intensities between neurons. W is the component to be learned to specify the relation between observations and actions. The REINFORCE algorithm is a learning rule of a particular form:

$$\Delta W_i = \alpha(r_t - b)e_i(t) \tag{1}$$

where α is a positive learning rate coefficient, b is a reinforce baseline, r_t is an immediate reward, and e_i is called the eligibility of W_i where W_i is each element in W.

The reward evaluates the agent's performance by taking an action. In general, a positive reward indicates that the action is preferred. In the REINFORCE learning rule, the update of weights is proportional to the expected reward increases.

The eligibility e_i specifies a correlation between weight and executed action and is presented as

$$e_i(t) = \frac{\partial}{\partial W_i} \ln\big(\pi(a_t, X_t, W)\big). \tag{2}$$

Rather than only considering the immediate reward at time *t*, the algorithm can record the history of the agent's performance. In this case, the algorithm is modified in such a way:

$$\Delta W_i = (r_t - b)D_i(t) \tag{3}$$

where $D_i(t) = e_i(t) + \gamma D_i(t-1)$ and $D_i(t)$ is a discounting running average of eligibility, and γ is a discount coefficient between zero and one.

It is interpreted that the update rule seeks the synaptic connection strengths to optimize the long-term average reward.

At each time t, the policy can be improved by updating W as

$$W = W + \alpha(1-\gamma)\Delta W \tag{4}$$

where $\Delta W = \begin{bmatrix} \Delta W_1 & \Delta W_2 & \cdots & \Delta W_i & \cdots \end{bmatrix}$

W is updated to increase the probability of actions according to the history when the averaged reward is positive. The update of weights is repeated until a performance criterion is achieved.

A general selection of stochastic policy distribution is the normal distribution when the action is in continuous space, because the normal distribution is a simple second-order sufficient statistics. This distribution requires the mean and the variance only. The two statistics can be regarded as the weights to be learned ($W = [\mu, \sigma]$).

Therefore, the stochastic policy has a form of

$$\pi(a_t, X_t, [\mu, \sigma]) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(a_t - \mu)^2}{2\sigma^2} \right) \tag{5}$$

The eligibilities of weights become $e_\mu = \frac{a_t - \mu}{\sigma^2}$ and $e_\sigma = \frac{(a_t - \mu)^2 - \sigma^2}{\sigma^3}$ accordingly. By setting learning rate coefficients appropriately, $e_\mu = a_t - \mu$ and $e_\sigma = \frac{(a_t - \mu)^2 - \sigma^2}{\sigma}$ can be obtained. This protects the eligibilities from divergence [Williams 1992; Kimura and Kobayashi 1998].

As pointed in Williams (1992), such a Gaussian unit is practically useful because the mean and variance of its output is individually adaptively controllable by using separate weights. Using multiparameter distributions enable to control their degrees of exploratory behavior.

From the perspective of the neural spiking activity, the action in discrete space was originally investigated in [Seung 2003; Bartlett and Baxter 2000]. Neural spikes can be interpreted as binary signals. Therefore, a more biologically plausible stochastic policy distribution is the Bernoulli distribution. It can be assumed that a neural spike evokes with the probability of p_t at each time t.

In summary, the online adaptive learning algorithm based on the REINFORCE with the policy history can be implemented to be a sequential procedure as follows.

1. Take observation inputs X_t from the environment at time t.

2. Take action a_t with probability $\pi(a_t, X_t, W)$.
3. The immediate reward r_t is evaluated.
4. Accumulating eligibility is calculated as $D_i(t) = e_i(t) + \gamma D_i(t-1)$ where $e_i(t) = \frac{\partial}{\partial W_i} \ln(\pi(a_t, X_t, W))$
5. Update the weights $W = W + \alpha(1-\gamma)\Delta W$ where $\Delta W = [\Delta W_1 \quad \Delta W_2 \quad \cdots \quad \Delta W_i \quad \cdots]$ with $\Delta W_i = (r_t - b)D_i(t)$.
6. Repeat the procedure at time $t+1$.

1.3. Underactuated Systems

A mechanical system with configuration vector $q \in Q$ and Lagrangian $L(q, \dot{q})$ satisfying the Euler-Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = F(q)u \tag{6}$$

is called an *Underactuated Mechanical System* if $m = \text{rank}F(q) < n = \dim(Q)$ where $F(q)$ maps control inputs into generalized forces . In other words, underactuated systems have fewer actuators than configuration variables. This restriction of the control authority does not allow exact feedback linearization of underactuated systems. In a special case satisfying $F(q) = [0 \quad I_m]^T$, the first $(m-n)$ equations in (6) represents the *unactuated subsystem* which is expressed as a second-order dynamic equation in the form of $\eta(\ddot{q}, \dot{q}, q) = 0$, and the remaining m equations the *actuated subsystem*. The actuated subsystem can be linearized using an invertible change of control while the unactuated subsystem remains as a nonlinear system. The procedure is called partial feedback linearization. The unactuated system dynamics is still coupled with the linearized actuated subsystem, which makes it difficult to design a controller for underactuated systems.

Interestingly, the control in many cases of human or animal movements can be regarded as the control of underactuated mechanical systems. In addition, a variety of biologically-inspired robotic designs such as passive biped walker, Acrobot or Pendubot etc are underactuated [Spong 1996; Spong 1999]. Therefore, the class of underactuated mechanical systems is rich. For example, suppose that a legged locomotion is analyzed with *N* internal joints and *N* actuators, the position and orientation of the body in space requires six degrees of freedom in addition. Therefore, *N* control inputs are fewer than the *N* + 6 degrees of freedom. Because most of the interesting problems of movements are underactuated, it is no surprise to try to investigate the underactuated mechanical system in order to understand control strategies of human or animal movements. Furthermore, the underactuated control systems may shed a light on producing more dynamic and more agile movement because classical fully-actuated control systems generally tend to augment control authority and

override the dynamics of the plant in such a way to strictly follow a desired trajectory while neglecting the natural dynamics. Recent control strategies tend to pursue exploiting the dynamics rather than canceling it. However, the nonlinear underactuated systems often include feedforward nonlinearities, unstable zero dynamics, and other structural properties that cause difficulties to apply some recent design techniques. A typical control design strategy of the underactuated mechanical system is to combine Lyapunov theory with passivity properties and energy shaping. However, a unique control design is hard to achieve the overall nonlinear performance at all; instead, switching control strategy is combined in general [Spong and Praly 1996;Spong 1996].

This chapter plans to explore a typical example of such underactuated system, the Cart-Pole system mentioned in previous section. Using the specific system, a biologically motivated learning control algorithm will be investigated.

2. Computational Model

2.1. Plant Model

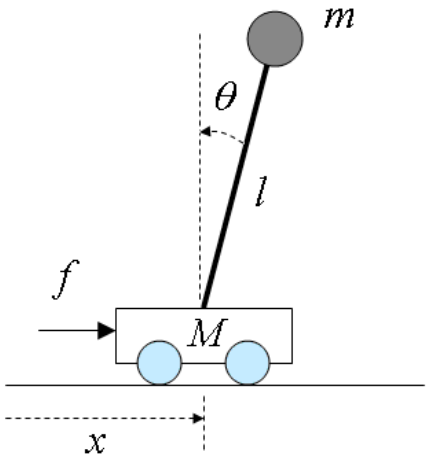

Figure 2. The cart-pole system.

The Cart-Pole system describes a structure where an inverted pendulum is hinged to a cart via a pivot and only the cart is actuated. Figure 2 illustrates the Cart-Pole system, where x is the horizontal position of the cart and θ is the clockwise angle between the pendulum and the vertical line. The goal of this task is to stabilize the pendulum in a vertical position and maintain the cart at original position ($\theta = 0$, $\dot{\theta} = 0$, $x = 0$, $\dot{x} = 0$). The pole is modeled to be a point mass attached to the top of a massless bar whose other end is pivoted to the cart. The mass of cart and pole is respectively selected to be 1 and 0.1 kg for simulation. The length of the bar is 1m. The cart-pole system moves horizontally on the flat ground under the effect of gravitational field.

The cart position is described by $\mathrm{x}_c = [x \quad 0]^T$ and the pendulum position is given by $\mathrm{x}_p = [x + l\sin\theta \quad l\cos\theta]^T$ with linear velocities $\dot{\mathrm{x}}_c = [\dot{x} \quad 0]^T$, and $\dot{\mathrm{x}}_p = [\dot{x} + l\dot{\theta}\cos\theta \quad -l\dot{\theta}\sin\theta]^T$.

The kinetic and potential energies of the system are, respectively,

$$T = \frac{1}{2}(m+M)\dot{x}^2 + m\dot{x}\dot{\theta}l\cos\theta + \frac{1}{2}ml^2\dot{\theta}^2$$

and

$$U = mgl\cos\theta .$$

Therefore, the Lagrangian of the Cart-Pole system can be expressed as

$$L = T - U = \frac{1}{2}\begin{pmatrix}\dot{x}\\ \dot{\theta}\end{pmatrix}^T \begin{bmatrix} m+M & ml\cos\theta \\ ml\cos\theta & ml^2 \end{bmatrix}\begin{pmatrix}\dot{x}\\ \dot{\theta}\end{pmatrix} - mgl\cos\theta . \tag{7}$$

The Euler-Lagrange equations of motion for the Cart-Pole system is in the form

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = f \text{ and } \frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}} - \frac{\partial L}{\partial \theta} = 0$$

which concludes the following equation of the motion:

$$(m+M)\ddot{x} + ml\ddot{\theta}\cos\theta - ml\dot{\theta}^2\sin\theta = f \tag{8}$$

$$ml\ddot{x}\cos\theta + ml^2\ddot{\theta} - mgl\sin\theta = 0 \tag{9}$$

A simple rearrangement yields in the standard form of:

$$\begin{bmatrix} m+M & ml\cos\theta \\ ml\cos\theta & ml^2 \end{bmatrix}\begin{pmatrix}\ddot{x}\\ \ddot{\theta}\end{pmatrix} + \begin{bmatrix} 0 & -ml\dot{\theta}\sin\theta \\ 0 & 0 \end{bmatrix}\begin{pmatrix}\dot{x}\\ \dot{\theta}\end{pmatrix} + \begin{bmatrix} 0 \\ -mgl\sin\theta \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{10}$$

Equation above represents clearly the Cart-Pole system belongs to the class of underactuated mechanical systems.

Though the system is not feedback linearizable completely, but a portion of the system can be linearized. The partial feedback linearization enables to be implemented on the dynamics of the actuated configuration variables. The procedure is called *collocated* partial feedback linearization [Spong 1996]. On the other hand, the partial feedback linearization procedure that linearizes the dynamics of the unactuated configuration variables is called *noncollocated* partial feedback linearization.

2.2. Control Algorithm

2.2.1. Swing-up Control

To design a controller of the cart-pole system, collocated partial feedback linearization is used. From equation (9), $\ddot{\theta}$ can be described with other terms

$$\ddot{\theta} = -\frac{1}{l}\left(\ddot{x}\cos\theta - g\sin\theta\right) \tag{11}$$

Plugging it into equation (8),

$$(m + M - m\cos^2\theta)\ddot{x} + mg\cos\theta\sin\theta - ml\dot{\theta}^2\sin\theta = f \tag{12}$$

To design a controller, the computed torque control method [Murray et al 1994; Craig 1986] is implemented. The computed torque control method is a popular approach in robotic control. When the system dynamics is described as $M(\theta)\ddot{\theta} + N(\theta,\dot{\theta})\dot{\theta} + G(\theta) = f$, the method sets a controller in a form of $f = \alpha f' + \beta$. Then, α and β are respectively chosen to be $\alpha = M(\theta)$, and $\beta = N(\theta,\dot{\theta})\dot{\theta} + G(\theta)$. In this way, the system equation can be simplified to be $\ddot{\theta} = f'$.

For the cart-pole system, let $f = (m + M - m\cos^2\theta)u + mg\cos\theta\sin\theta - ml\dot{\theta}^2\sin\theta$, then,

$$\ddot{x} = u \tag{13}$$

$$\ddot{x} = u$$

From equation (11), $$\ddot{\theta} = -\frac{1}{l}\left(u\cos\theta - g\sin\theta\right). \tag{14}$$

Now u should be designed. In the classical control scheme, $u = \ddot{\theta}_d + k_v(\dot{\theta}_d - \dot{\theta}) + k_p(\theta_d - \theta)$, which forces joint motion to follow desired trajectory.

Another potential scheme is passivity-based control strategy. The energy of the pendulum is

$$E_p = \frac{1}{2}ml^2\dot{\theta}^2 + mgl\cos\theta. \tag{15}$$

The target energy when the pendulum stays upward vertically is

$$E_d = mgl. \tag{16}$$

Using the energy difference of $\tilde{E} = E_p - E_d$, it is easily driven that

$$\dot{E} = ml^2\dot{\theta}\ddot{\theta} - mgl\dot{\theta}\sin\theta = -ml^2\dot{\theta}\left(\frac{u\cos\theta - g\sin\theta}{l}\right) - mgl\dot{\theta}\sin\theta = -mlu\dot{\theta}\cos\theta$$

To assure the dissipation of the energy, u can be chosen to be $u = \frac{k_s}{ml}\dot{\theta}\cos\theta\tilde{E}$ because this yields $\dot{\tilde{E}} + k_s\left(\dot{\theta}\cos\theta\right)^2\tilde{E} = 0$ therefore $\tilde{E} \to 0$ exponentially as time goes. In addition, for the regulation of x feedback component is designed. While the pendulum swings up to pass through the upright position, the cart position is also controlled to stay at origin. Only considering the cart's horizontal motion, a time-varying surface in motion space is described by a scalar equation $S(x;t) = \ddot{\tilde{x}} + p_1\dot{\tilde{x}} + p_2\tilde{x} = \ddot{x} + p_1\dot{x} + p_2 x = 0$ where $\tilde{x} = x - x_d$ and $\ddot{x}_d = \dot{x}_d = x_d = 0$. The cart behavior is expected to reach the surface within a finite time and stay there. To attain the goal, a controller of $u = -p_1\dot{x} - p_2 x$ is selected with equation (13). Neglecting the pendulum motion, a smooth cart motion could be described by a second order system with a natural frequency and a damping ratio. Setting the damping ratio to 1, $p_2 = \left(\frac{p_1}{2}\right)^2$ is applicable. p_2 is approximately a half of the natural frequency. To attain a stable and smooth cart motion, it is good that p_2 is less than 1.

As a result, the following controller form is designed to be:

$$u = \frac{k_s}{ml}\dot{\theta}\cos\theta\tilde{E} - \frac{k_c^2}{4}x - k_c\dot{x}. \tag{17}$$

The controller is guaranteed to converge on the orbit where $\tilde{E} = 0$. This energy-based control design of the Cart-Pole system has been investigated by many researchers [Spong 1999; Chung and Hauser 1995]. The status assures the pendulum passes through the upright position. But, the guaranteed stability is not asymptotic to a point but to an orbit. For this reason, the control strategy, in general, has to be switched to achieve an upright balance. In fact, in a neighborhood of the upright position, more precisely speaking, within the basin of attraction of the balancing controller, a linear feedback balancing controller is sufficient to achieve a fixed equilibrium point, which is a vertical posture, so that such switching control design is very simple and widely applicable. However, the difficulty of the switching control is to determine conditions on when to switch controllers. In addition, proper selection of the control gains such as k_s and k_c is important to achieve desired performance regardless of initial conditions. In most cases, lots of trial and error experiments are required to decide the switching conditions and control gain values. This chapter tests the effectiveness of the stochastic reinforce algorithm as an online control strategy that automatically satisfies the two requirements, switching condition and gain setup.

2.2.2. Balancing Control

Because the energy based swing-up control cannot maintain a upright balance of the pendulum, the controller is switched when the pendulum reaches within a certain range of space.In the basin of attraction of the balancing controller, the balancing controller is designed simply to be a linear feedback controller as follows.

From equation (14), the dynamics of pitch angle within a basin of attraction of a balancing controller, which is being designed, is described as

$$\ddot{\theta} = \frac{g}{l}\sin\theta - \frac{1}{l}u\cos\theta$$

Furthermore, it behaves quite linearly near the vertical balancing position such as

$$\ddot{\theta} = \frac{g}{l}\theta - \frac{1}{l}u \tag{18}$$

because $\tan\theta \cong \theta$, $\cos\theta \cong 1$ for θ is small.

Therefore, a switching condition is determined by the feasibility of the linear approximation above. For the current test, $|\theta| < \theta_C$ is set as the switching criterion where θ_C is a constant. A control law in equation (18) is pursued to maintain the pendulum vertically. To achieve a critical damping behavior with no oscillations only with respect to the pendulum dynamics, let

$$u = -(1 + k_b l)g\theta - 2l\sqrt{k_b g}\dot{\theta}. \tag{19}$$

With the above choice of the control law, the system becomes $\ddot{\theta} + 2\sqrt{k_p}\dot{\theta} + k_p\theta = 0$. Therefore, $\theta \to 0$ as time goes on.

In addition, regulation of the cart position is also added.

Therefore,

$$u = -(1 + k_b l)g\theta - 2l\sqrt{k_b g}\dot{\theta} - k_{p2}x - k_{v2}\dot{x} \tag{20}$$

2.3. Switching Condition

As mentioned previously, balancing control operates while $|\theta| < \theta_C$, and swing-up control operates while $|\theta| \geq \theta_C$. To confirm the switching condition, θ_C should be determined. The threshold value indicates a feasible angle range where balancing control is applied.

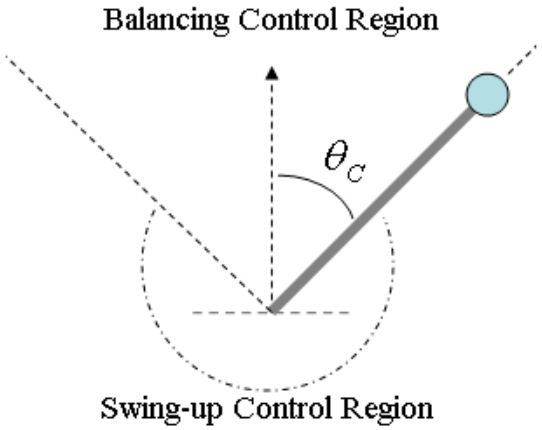

Figure 3. Two control strategies and switching condition.

It is proposed that θ_C is computed at the end of each trial as

$$\theta_C = \begin{cases} \min(\theta) & if\ \min(\theta) < 1.4 \\ 0.6 & if\ \min(\theta) > 1.4 \end{cases} \tag{21}$$

When the swing-up control does not reach the right track yet for initial trials, the threshold is fixed to 0.6. Once the swing-up control functions effectively, the threshold value is updated according to equation (21). Principally when the angular position reaches at highest point, the balancing control operates. If the switching position is not appropriate or the balancing control is not well tuned, the trial will fail, but control parameters will be updated by a learning algorithm as explained in next section. In next trial, updated control parameters will be used. If the switching position is within the range such that the balancing control can effectively grasp the pendulum straight up, the test will succeed, and control parameters and switching condition are confirmed.

2.4. Learning Algorithm

Using the online stochastic learning algorithm explained earlier, the control gains k_s, k_{p1}, k_{v1}, k_{p2}, k_{v2} and k_b are found once a set of physical parameters of the system is given.

Now, the mean and standard deviation of the stochastic policy are defined respectively as

$$\mu = -(1+W_1 l)g\theta_t - 2l\sqrt{W_1 g}\dot{\theta}_t - W_2 x_t - W_3 \dot{x}_t, \text{ and } \sigma = 0.1 + \frac{1}{1+\exp(-W_7)} \tag{22}$$

when the switching condition is satisfied, i.e., $|\theta_t| < \theta_C$, and

$$\mu = W_4 \tilde{E} \dot{\theta}_t \cos\theta_t - \frac{W_5^2}{4} x_t - W_5 \dot{x}_t \text{, and } \sigma = 0.1 + \frac{1}{1+\exp(-W_6)} \tag{23}$$

when switching condition is not satisfied, i.e., $|\theta_t| \geq \theta_C$.

Then, eligibilities are given by $e = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{bmatrix}^T$, where

$$e_1 = -(a_t - \mu) l \left(g\theta_t - \sqrt{\frac{g}{W_1}} \dot{\theta}_t \right),\ e_2 = -(a_t - \mu) x_t,\ e_3 = -(a_t - \mu)\dot{x}_t,\ e_4 = 0,\ e_5 = 0,$$
$$e_6 = \left((a_t - \mu)^2 - \sigma^2 \right)(1 + 0.1 - \sigma) \text{ when } |\theta_t| < \theta_C,$$

and

$$e_1 = 0,\ e_2 = 0,\ e_3 = 0,\ e_4 = (a_t - \mu)\tilde{E}\dot{\theta}_t \cos\theta_t,\ e_5 = -(a_t - \mu)\left(\frac{W_5}{2} x_t + \dot{x}_t \right),$$
$$e_6 = \left((a_t - \mu)^2 - \sigma^2 \right)(1 + 0.1 - \sigma) \text{ when } |\theta_t| \geq \theta_C.$$

Considering the above eligibilities with the weight update rule in equation (3), it is worthwhile to recognize that the overall learning rule contains a Hebbian component when the observed information is regarded as inputs and reward as outputs.

The control gain is updated by equation (4) and the process is repeated until a target performance of the system is satisfied. The next step is to define a reward. The reward baseline is set to zero ($b = 0$) in this test.

If the cart is way out of the desired position in 30 seconds after the simulation begins, the simulation is terminated and an immediate reward is assigned.

$$r_t = -5 \text{ if } |x_t| > 3 \text{ or } |\theta_t| > \pi/6 \text{ in 30 seconds} \tag{24}$$

Or the simulation is implemented until a trial ends. A trial takes 60 seconds. After a trial, the following reward is assigned (Figure 4).

$$r_t = +10 \text{ if } |x_t| < 0.1 \text{ and } |\theta_t| < \pi/9, \tag{25}$$
$$r_t = -5 \text{ otherwise.}$$

When a task is successful, the learning process stops, and control gain values and switching conditional parameter are confirmed. The success of a task is computationally accepted if conditions are satisfied such that $|\theta(T_f)| < 0.05$, $|x(T_f)| < 0.05$, $|\dot{\theta}(T_f)| < 0.005$, and $|\dot{x}(T_f)| < 0.005$, where T_f indicates a final time. The conditions guarantee a convergence to upright posture of the pendulum.

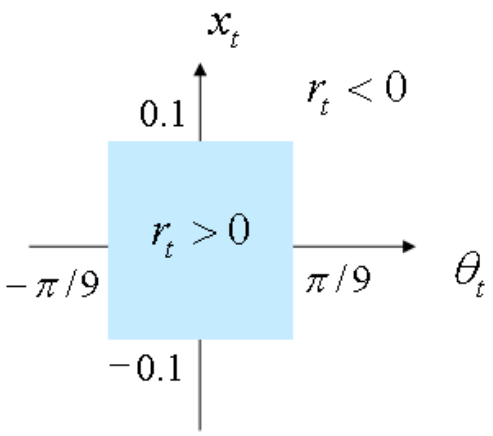

Figure 4. Immediate reward function.

3. Results

The simulation task was implemented to test the proposed algorithm. The initial position of the cart was at the origin ($x(0) = 0$) and the initial position of the pendulum was almost vertical, pointing downwards with no velocity ($\theta(0) = 3.12$). Control gains were randomly initialized. In the simulation, physical parameters were set to be $m = 0.1$, $M = 1, l = 1$ and $g = 9.8$.

Figure 5 demonstrates a task simulation after obtaining control gains from learning tasks. For this task, initial control gain values were randomly chosen to be $W(0) = [0.6813 \quad -0.3795 \quad -0.8318 \quad 0.5028 \quad 0.7095 \quad 0.4289]^T$.

After 553 trials, the algorithm achieved a successful task. In Figure 5, each motion snapshot was recorded in every 0.2 second. At the beginning, the pendulum sway became greater and greater until the pendulum reached a certain height in about 25 seconds. Then, the cart moved quickly to evoke a reactive response of the pendulum, standing up and returns smoothly near the origin. Finally, balancing control leaded the positions of both the cart and pendulum to the desired ones. Control gains and switching threshold obtained from learning were:

$$W = [1.1736 \quad -0.3092 \quad -0.8027 \quad 0.3563 \quad 0.2949 \quad 1.1945]^T \text{ and } \theta_C = 0.9035$$

Depending on the initial control gain sets, convergence to a successful task required different numbers of trials during learning.

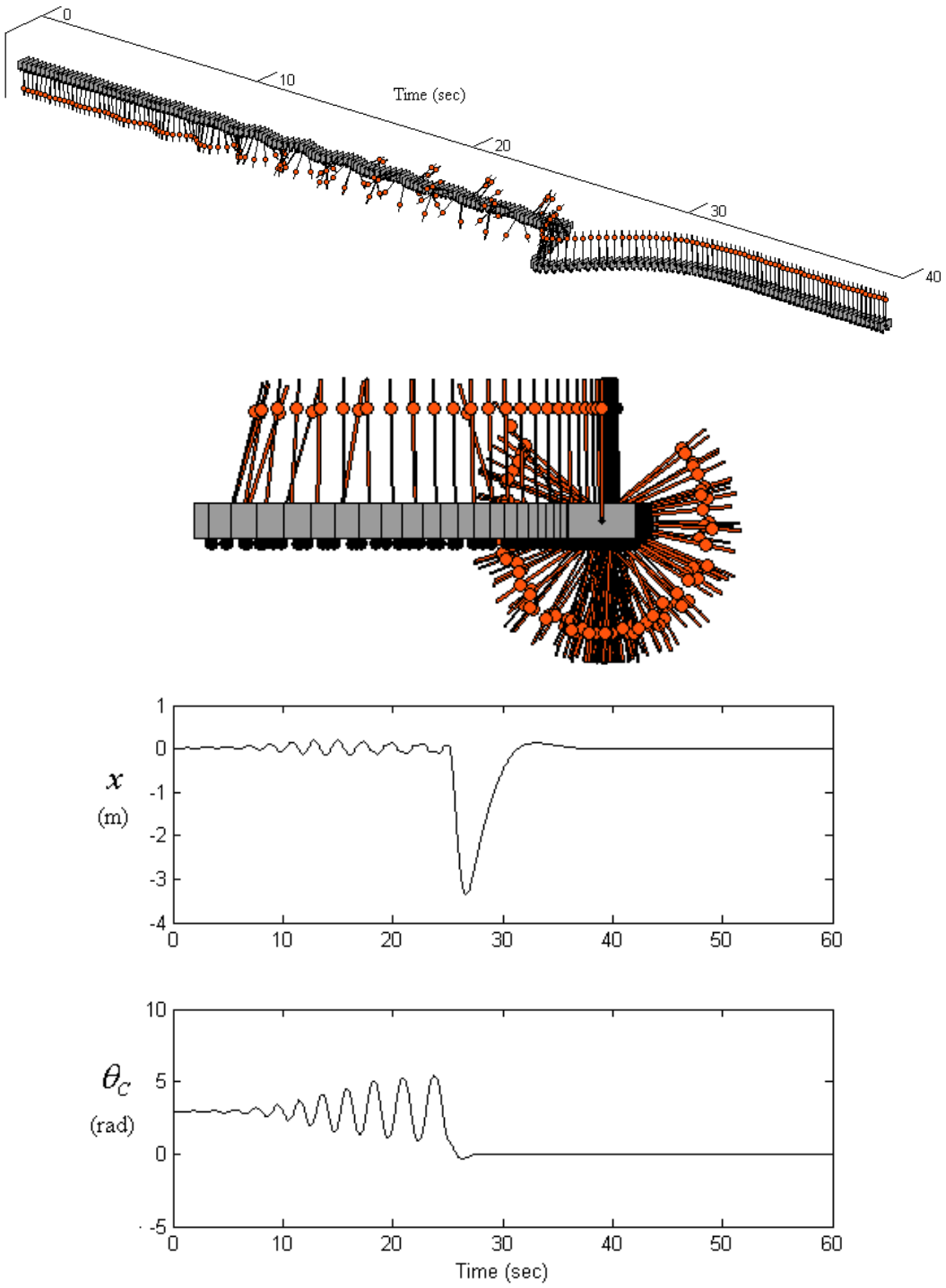

Figure 5. A task simulation (top): motion snapshots as a function of time, (middle): motion drawn in x-y coordination, (bottom): plots of cart and pendulum position trajectories.

4. Neurobiological Interpretation

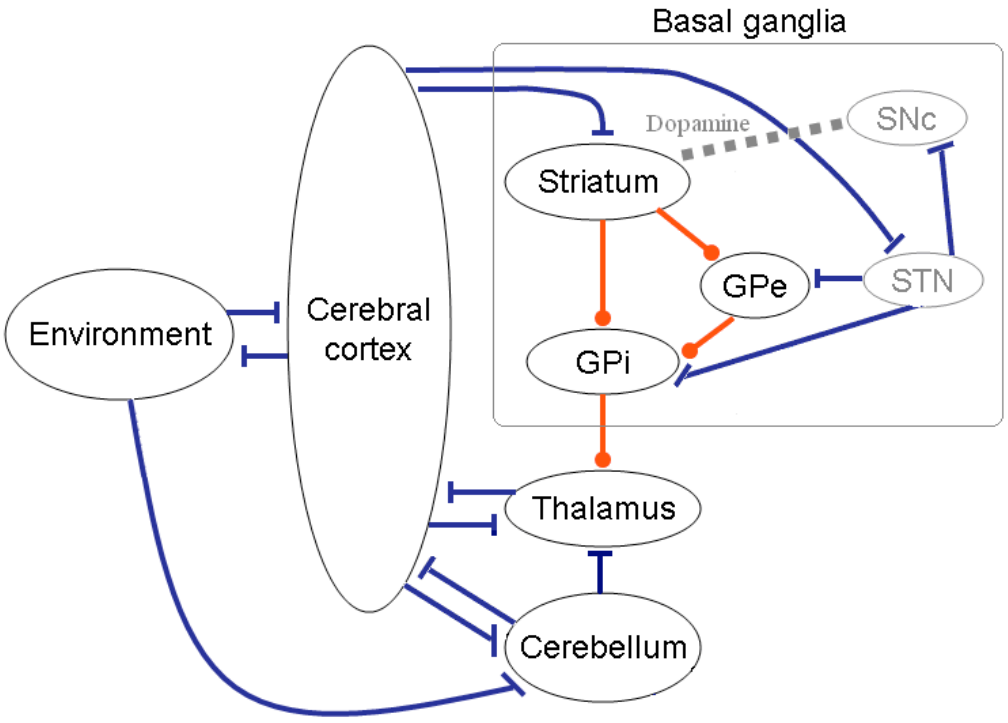

Figure 6. An illustration of neural signal processing between cerebrum, cerebellum, basal ganglia, and environment.

Figure 6 illustrates plausible neural connections to explain the cart-pole learning and control process in terms of neural signal processing.

From the outcome, the control signals are expressed in the form of

$$\mu_1 = -21.5\theta_t - 6.84\dot{\theta}_t + 0.31x_t + 0.8\dot{x}_t$$

$$\mu_2 = 0.36\tilde{E}\dot{\theta}_t \cos\theta_t - 0.02x_t - 0.29\dot{x}_t$$

According to parallel signal processing in the BG, it can be regarded that two control signals operate separately.

Based on previous statement in section 1.1, each control signal is interpreted as an effective sum of activities from direct and indirect pathways. In each control signal, positive coefficients can be interpreted as excitatory signal processing as the direct pathway, and negative coefficients as inhibitory signal processing along the indirect pathway.

The reinforce algorithm, a learning process of tuning control gains to take an appropriate action, may correspond to long-term plasticity in the striatum. The reinforce algorithm is online model-free learning process. It is recognized that phasic increase or decrease of dopamine concentration with the corticostriatal synaptic activity causes either synaptic long-term potentiation (LTP) or depression (LTD) [Wickens 2000;Schultz 1995; Mao 2005].

Positive and negative reward values could be comparable with the increase or decrease of dopamine concentration respectively. Increase and decrease of a control gain value are also comparable with LTP and LTD respectively. Doya et al. proposed the reinforcement learning model of the BG [Doya et al 2001; Doya 1999]. A model-free TD error learning model of Doya's is principally similar to the model in this article.

As the cerebellum receives rich input from many parts of the body, and Purkinje cell activity in the cerebellar cortex has been related to joint kinematic activity [Johnson and Ebner 2000], and it is highly conceivable that different corticonuclear microcomplexes become active in different kinematic states. Therefore, the scheduling of control signal according to sensed kinematic state is biologically plausible. This article focuses on control and learning process in BG. However, the learning algorithm requires a switching mechanism. A possible neural model of the switching (scheduling) mechanism can be adapted from Jo and Massaquoi [2004] as in Figure 7.

The cerebellar microzone has uniform neural circuits. It receives two types of inputs, one through mossy fibers, and the other through climbing fiber. Mossy fiber inputs excite Purkinje cells, which haves fan-like dentrites, propagating along parallel fibers. Each Purkinje cell inhibits deep cerebellar nuclear (DCN) cells. Parallel fibers inhibit Purkinje cells in neighborhood via basket cells while they excite Purkinje cells located in parallel with the parallel fiber beams. Based on the neuroanatomy, a simple proposal for the selection mechanism in the cerebellar cortex in which a beam of activity on "suppress" parallel fibers ($\mathrm{PF}_{\mathrm{sup}}$) inhibits Purkinje cells within a certain range [Eccles et al 1967; Ito 1984] (Figure 7). This diminishes the net inhibition in those modules, allowing them to process the signal that arrives on "signal" parallel fibers ($\mathrm{PF}_{\mathrm{sig}}$). Conversely, the beam activates local Purkinje cells, thereby suppressing the activity of the modules. The principal characteristic required of $\mathrm{PF}_{\mathrm{sup}}$ fibers in this scheme is that unlike $\mathrm{PF}_{\mathrm{sig}}$ fibers, they should contact Purkinje cells relatively more strongly than the corresponding DCN cells – if they contact the same DCN cells at all. This appears to be generally consistent with the experimental studies [Eccles et al 1974a; Eccles et al 1974b; Ito 1984]. A prime candidate source for suppressor parallel fibers is the dorsal spinocerebellar tract (DSCT) elements of which are known to convey mixtures of proprioceptive and other information from multiple muscles within a limb [Oscarsson 1965; Bloedel and Courville 1981; Osborn and Poppele 1992] while typically maintaining a steady level of background firing in the absence of afferent input [Mann 1973]. The observations are possibly formalized by proposing that the DSCT fibers transmit switching criterion: $|\theta| < \theta_C$ or $|\theta| \geq \theta_C$. The sensed state θ would correspond to the average signal of a large number of primary afferents (Mann 1973) and θ_C is assumed to be provided from the cerebral cortex. Thus, certain suppressor fibers become relatively more active when the sensed kinematic state is located in a region of the state space, i.e., the switching criterion is satisfied. The net switching action can therefore be written:

$$\mu = \mu_1 \left\lfloor 1 - \left[1 - \gamma \left\lfloor |\theta| - \theta_C \right\rfloor_+ \right]_+ \right\rfloor_+ + \mu_2 \left\lfloor 1 - \left[1 - \gamma \left[\theta_C - |\theta| \right]_+ \right]_+ \right\rfloor_+$$

where γ represents the strength of lateral inhibition provided by basket cells. This parameter regulates the steepness of the transition zone between different regions.

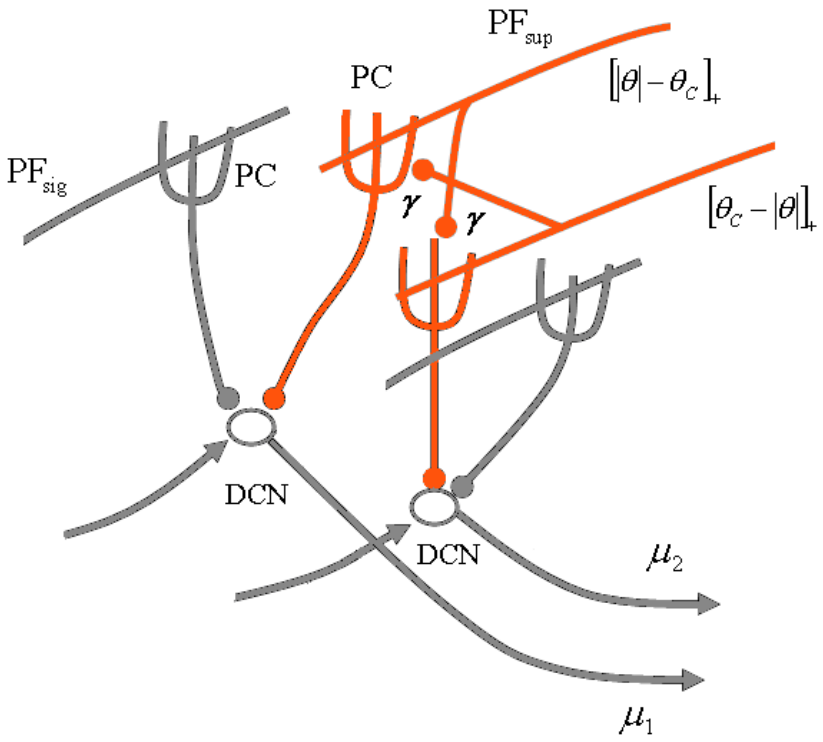

Figure 7. A proposal of selection mechanism in the cerebellar cortex, PC: Purkinje cell, DCN: deep cerebellar nuclei, PF: parallel fibers, each filled circle represents inhibition.

5. Conclusion

This chapter provided a computational demonstration of a learning algorithm. This algorithm was based on stochastic reinforcement learning combined with a switching mechanism. The algorithm performed a successful learning for the control problem of the Cart-Pole system. The learning algorithm was model-free and real-time adaptive. In addition, neuro-computational interpretation was given to explain each component in the learning algorithm. Especially, the control signal processing and reward-based update rule can be interpreted in terms of basal ganglionic neural computations, while the switching between swing-up and balancing controls can be explained with a plausible neural operation in the cerebellum.

References

Bartlett PL, Baxter J (2000) *A biologically plausible and locally optimal learning algorithm for spiking neurons*, Technical Report, Australian National University. http://arp.anu.edu.au/ftp/papers/jon/brains.pdf.gz

Kimura H, Kobayashi S (1998) Reinforcement learning for continuous action using stochastic gradient ascent, *Intelligent Autonomous Systems* (IAS-5).

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine Learning* **8**, 229-256.

Florian RV (2005) A reinforcement learning algorithm for spiking neural networks, *Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (SYNASC 05) 299-306.

Seung HS (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission, *Neuron* **40**(6), 1063-1073.

Murray RM, Li Z, Sastry SS (1994) *A mathematical introduction to robotic manipulation*, CRC Press, Inc.

Spong MW, Praly L (1996) Control of underactuated mechanical systems unsing switching and saturation, *Proceedings of the Block Island Workshop on control using logic based switching.*

Spong MW (1996) Energy based control of a class of underactuated mechanical systems, 1996 *IFAC World Congress*, July.

Spong MW (1999) Passivity-based control of the compass gait biped, *IFAC World Congress*, Beijing China, July.

Chung CC, Hauser J (1995) Nonlienar control of a swinging pendulum, *Automatica* **31**(6), 851-862.

Craig JJ (1980) *Introduction to Robotics Mechanics and Control*, 2nd Ed., Addison-Wesley Publishing Company.

Florentin W, Porr B (2005) Temporal sequence learning, prediction and control - A review of different models and their relation to biological mechanisms, *Neural Computation,* **17**, 245-319.

Berns GS, Sejnowski (1994) A model of basal ganglia function unifying reinforcement learning and action selection, *Proceedings of the joint symposium on neural computation*, La Joila, CA, 129-145.

Olfati-Saber R (1999) Fixed point controllers and stabilization of the cart-pole system and the rotating pendulum, *IEEE 38th Conference on Decision and Control*, Phoenix, AZ, Dec, 1174-1181.

Karr CL (1991) Design of a cart-pole balancing fuzzy logic controller using a genetic algorithm, *Proceedings of SPIE conference on Applications of Artificial Intelligence IX*.

Saravanan N, Fogel DB (1995) Evolving neural control systems, *IEEE Expert Magazine* **10**(3), 23-27.

Mao Z-H (2005) *Modeling the role of the basal ganglia in motor control and motor programming*, PhD thesis, Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology.

Doya K, Kimura H, Kawato M (2001) Neural mechanisms of learning and control, *IEEE control systems Magazine*, **42-54**.

Doya K (1999) What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural Networks*, 12.

Osborn CE, Poppele RE (1992) Parallel distributed network characteristics of the DSCT, *J Neurophysiol* **68**(4): 1100-1112.

Oscarsson O (1965) Functional organization of the spino- and cuneocerebellar tracts, *Phys Rev* **45**: 495-522.

Mann MD (1973) Clarke's column and the dorsal spinocerebellar tract: A review, *Brain Behav Evol* **7:** 34-83.

Ito M. (1984) *The Cerebellum and Neural Control*, New York, Raven Press.

Eccles JC et al. (1967) *The cerebellum as a neuronal machine*, New York, Springer-Verlag.

Eccles JC et al. (1974a) Temporal patterns of responses of interpositus neurons to peripheral afferent stimulation, *J Neurophysiol* **37**: 1427-1437.

Eccles JC et al. (1974b) Patterns of convergence onto interpositus neurons from peripheral afferents, *J Neurophysiol* **37**: 1438-1448.

Bloedel JR, Courville J (1981) Cerebellar afferent systems, In: Bookhart JM, Mountcastle VB, Brooks VB and Geiger SR (Eds), *Handbook of physiology: the nervous system II*, Bethesda, MD, American Physiological Society, Section I, Volume II: Motor Control, Part 2: 735-829.

Johnson MTV, Ebner TJ (2000) Processing of multiple kinematic signals in the cerebellum and motor cortices, *Brain Res Rev* **33**: 155-168.

Jo S, Massaquoi SG (2004) A Model of Cerebellum Stabilized and Scheduled Hybrid Long-loop Control of Upright Balance, *Biol Cybern* **91**(3):188-202.

Schultz W et al (1995) Reward-related signals carried by dopamine neurons, In: Houk JC, Davis JL, Beiser DG (eds.), *Models of Information Processing in the Basal Ganglia*, 233-248, MIT Press, Cambridge, MA.

Mink JW (1996) The basal ganglia: focused selection and inhibition of completing motor programs, *Progress in Neurobiology*, **50**:381-425.

Nolte J (1999) *The human brain: an introduction to its functional anatomy (4th ed.)*, Mosby-Year Book, Boston, MA.

Wickens JR (2000) Dopamine regulation of synaptic plasticity in the neostriatum: a cellular model of reinforcement, In: Miller R, Wickens JR (Eds), *Brain dynamics and the striatal complex*, **65-76**, Harwood Academic Publishers, Australia.

Kelly RM, Strick PL (2004) Macro-achitecture of basal ganglia loops with the cerebral cortex: use of rabies virus to reveal multisynaptic circuits, *Progress in Brain Research*, **143**: 449-459.