Q Learning based Reinforcement Learning Approach to Bipedal Walking Control

Sudhir Raj
Mechanical Engineering Department
IIT Kharagpur
Kharagpur, India
E-mail: 10ME90R26@iitkgp.ac.in

Cheruvu Siva Kumar
Mechanical Engineering Department
IIT Kharagpur
Kharagpur, India
E-mail: kumar@mech.iitkgp.ernet.in

***Abstract*—Reinforcement learning has been active research area not only in machine learning but also in control engineering, operation research and robotics in recent years. It is a model free learning control method that can solve Markov decision problems. Q-learning is an incremental dynamic programming procedure that determines the optimal policy in a step-by-step manner. It is an online procedure for learning the optimal policy through experience gained solely on the basis of samples. A Q learning based reinforcement learning of a double inverted pendulum has been shown in this paper which reaches a limit cycle at the end of several learning cycles. The double inverted pendulum becomes stable, since the pole angle and pole angular velocity become zero. Stabilization of an equivalent double inverted pendulum representing a bipedal robot has been successfully implemented for balancing the pole angles in the required range using Q learning in Reinforcement Learning.**

Keywords—Q learning; Double inverted pendulum; Limit Cycle.

I. INTRODUCTION

With advances in science and technology, the interest to study the human walking has developed the demand for building the humanoid robot. By making the robot fully autonomous, it can be used in the environments where humans cannot enter. Complex movements can be achieved by increasing the degrees of freedom. There are several methods of designing a stable walking gait pattern. The first approach is the inverted pendulum model control method. Another new approach is to use a neuro dynamic controller. Q learning can be used to find an optimal selection policy for any given Markov decision process. It works by learning an action value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter. The reinforcement learning method uses the Q learning algorithm, which uses the Q value. The humanoid robot is modeled as double inverted pendulum.

A lot of institutes and researchers are actively working towards development of humanoid robot and efforts are given to develop the improved control system for humanoid robot.

Zho et al. [1] describe Vague Neural Network based control system which is implemented to balance a cart pole system. A new reinforcement learning algorithm of neural network is proposed. Simulation results of inverted pendulum show that the two output neurons play different roles in reinforcement learning, the combination of them has an excellent effect on the Q learning result.

Bogdanov [2] describes optimal control of a double inverted pendulum on a cart. Problem of optimal control minimizing a quadratic cost function is addressed. Linear quadratic regulator (LQR), State dependent ricatti equation (SDRE), optimal neural network (NN) control and combination of the NN with the LQR and SDRE has been tested.

Kaynov [3] describes about open motion control architecture for humanoid robot. This thesis proposes joint motion control problem and a new solution to walking stability problem for humanoids. A new original walking stabilization controller based on decoupled inverted pendulum dynamic model is developed.

Shaoqiang et al. [4] describe modelling and simulation of robot based on Matlab/SimMechanics. The SimMechanics block model is first used in modelling and simulation of inverted pendulum. Simulation results of the SimMechanics Block model and mathematical model for single inverted pendulum is compared. A full state feedback controller is designed to satisfy the performance requirement.

Park et al. [5] describe stabilization of biped robot based on two mode Q learning. Two mode Q learning, an extension of Q learning is used to stabilize the zero moment point of a biped robot in the standing posture. In the two mode Q learning, the experiences of both success and failure of an agent are used for fast convergence. The effectiveness of two mode Q learning is verified by the use of real experiment.

Suleiman et al. [6] describe enhancing zero moment point based control model which uses system identification approach. The approximation of a humanoid robot by an inverted pendulum is one of the most frequently used models to generate a stable walking pattern using a planned zero moment point trajectory. The accuracy of the inverted pendulum using system identification techniques has been proposed.

Kim et al. [7] describe ZMP based neural network inspired humanoid robot control. To ensure a steady and smooth walking gait of such robots, a feedforward type of

neural network architecture trained by the back-propagation algorithm is employed.

Asano et al. [8] describe passive dynamic walking which is used as a reference model. The control design technique used in this study was shown to be effective in generating a walking pattern, and its validity has been proved by numerical simulations and experiments.

Oh et al. [9] describe an analytic method to generate the real time trajectory of the center of mass is proposed for given zero moment point pattern. The whole walking process is divided into transient and periodic walking phases. For each phase of walking, the analytic solution of the center of mass for the given ZMP based on the inverted pendulum model is computed.

Tang and Er [10] describe a planning method for humanoid walking. Inverted pendulum model (IPM) is used as a dynamic model for humanoid robots. Zero moment point constraints of the robot are analyzed in the IPM motion, and the COG (center of gravity) motion of IPM is to approximate the COG motion of robots.

Thant et al. [11] present cubic spline interpolation based trajectory planning method which is aiming to achieve smooth biped robot walking trajectory. The walking trajectory of bipedal robot has been achieved using cubic spline interpolation.

Gullapalli et al. [12] present Stochastic real-valued (SRV) reinforcement learning algorithm, and it is used for learning control, and it can be used with nonlinear (multilayer) artificial networks. In the peg-in-hole insertion task, SRV network successfully learns to insert a peg into a hole with extremely low clearance, in spite of high sensor noise.

Schaal et al. [13] present a probabilistic reinforcement learning approach, which is derived from the framework of stochastic optimal control and path integrals. The policy improvement with path integrals (PI^2) is able to efficiently learn humanoid motor skills which require full-body motion and variable impedance control, and involve direct contact with the environment.

Vijaykumar et al. [14] present the design, construction and preliminary testing of a planar bipedal robot with joints capable of physically varying both their stiffness and damping independently- the first of its kind. A wide variety of candidate variable stiffness and damping actuator designs are investigated.

Morimoto et al. [15] present a method for learning biped locomotion from demonstration and its frequency adaptation using dynamical movement primitives. Demonstrated trajectories are learned through movement primitives by locally weighted regression, and the frequency of the learned trajectories is adjusted automatically by a frequency adaptation algorithm based on phase resetting and entrainment of coupled oscillators.

Franklin et al. [16] present biped dynamic walking using reinforcement learning. The self scaling Reinforcement learning algorithm was developed in order to deal with the problem of reinforcement learning in continuous action domains.

Stabilization of double inverted pendulum using Q learning in Reinforcement learning was not addressed in the previous work. Earlier models did not consider stiffness and damping of joints. However in practice the motors and gear boxes used have their characteristics in terms of stiffness, damping and/or friction. Using these in model is also considered. The motivation of this work is to develop neuro dynamic control in bi-pedal walking.

II. MODELLING OF HUMANOID ROBOT

Humanoid bipedal walking is often considered in two planes by most researcher [1] as Frontal plane biped motion and sagittal plane motion. Both of these have various phases and the important dynamical conditions can be considered in a periodic manner. A Complete walking cycle is composed of two phases. These two phases are double support phase and single support phase. During the double support phase, both feet are in contact with the ground. During the single support phase, while one foot is stationary with the ground, the other foot swings from the rear foot to the front and is shown in figure 1.

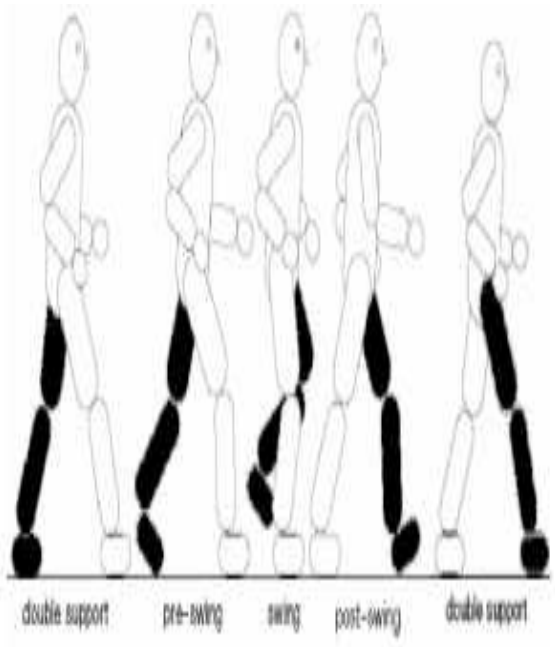

Fig. 1 Walking Phase of humanoid robot taken from [11]

Walking alternates between a double support phase and single support phase. The single support phase faces the maximum variation of dynamic/body weight transfer/transition during the walking. It is of most interest for researchers. This is initially modelled as double inverted pendulum [3]. Humanoid robot can be considered as double inverted pendulum during the single support phase.

As an initial study for simplified motion dynamics of a leg and the humanoid, a double inverted pendulum equivalent is considered. Figure 2 shows model of such a double inverted pendulum.

The equivalences are as follows:

1. Body mass is at m_2. Body is assumed to be rigid.
2. m_1 is the equivalent mass at knee.
3. Foot is fixed at 'O' during the contact part of motion.

k_1 and c_1 are stiffness and damping coefficients of joint 1.

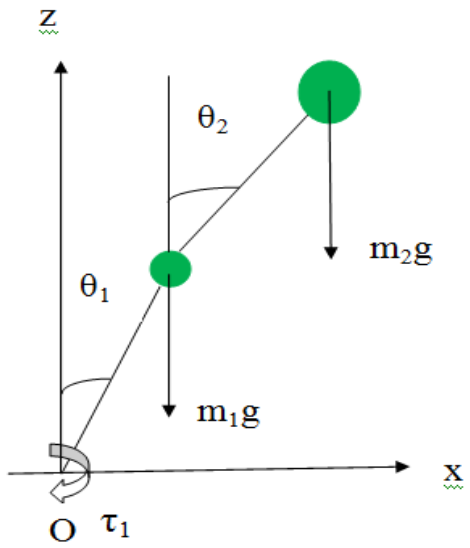

Fig. 2 Double inverted pendulum

Where
l_1 = length of first rigid massless link
l_2 = length of second rigid massless link
m_1 = mass of the first pendulum
m_2 = mass of the second pendulum
θ_1 = angle of the first link with vertical
θ_2 = angle of the second link with vertical

$$x_1 = l_1 \sin\theta_1 \quad (1)$$
$$z_1 = l_1 \cos\theta_1 \quad (2)$$
$$x_2 = l_1 \sin\theta_1 + l_2 \sin\theta_2 \quad (3)$$
$$z_2 = l_1 \cos\theta_1 + l_2 \cos\theta_2 \quad (4)$$
$$V = m_1 g z_1 + m_2 g z_2 \quad (5)$$
$$V = (m_1 + m_2) g l_1 \cos\theta_1 + m_2 g l_2 \cos\theta_2 \quad (6)$$
$$T = \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \quad (7)$$
$$v_1^2 = \dot{x}_1^2 + \dot{z}_1^2 \quad (8)$$
$$v_2^2 = \dot{x}_2^2 + \dot{z}_2^2 \quad (9)$$

$$T = \frac{1}{2} m_1 l_1 \dot{\theta}_1^2 + \frac{1}{2} m_2 [l_1^2 \dot{\theta}_1^2 + l_2 \dot{\theta}_2^2 + 2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2)] \quad (10)$$

$$L = T - V \quad (11)$$

$$L = \frac{1}{2}(m_1 + m_2) l_1^2 \dot{\theta}_1^2 + \frac{1}{2} m_2 l_2^2 \dot{\theta}_2^2 + m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2) - (m_1 + m_2) g l_1 \cos\theta_1 - m_2 g l_2 \cos\theta_2 \quad (12)$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_i}\right) - \frac{\partial L}{\partial \theta_i} = \tau_i \quad (13)$$

$$(m_1 + m_2) l_1^2 \ddot{\theta}_1 + m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) - l_1 g (m_1 + m_2) \sin\theta_1 = \tau_1 - k_1\theta_1 - c_1\dot{\theta}_1 \quad (14)$$

$$m_2 l_2^2 \ddot{\theta}_2 + m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) - m_2 l_1 l_2 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) - l_2 m_2 g \sin\theta_2 = 0 \quad (15)$$

III. Q LEARNING CONTROL

Q learning is a recent form of Reinforcement Learning algorithm that does not need a model of its environment and can be used on-line. Q learning algorithms works by estimating the values of state-action pairs. The value Q(s, a) is defined to be the expected discounted sum of future payoffs obtained by taking action a from state s and following an optimal policy thereafter. Once these values have been learned, the optimal action from any state is the one with the highest Q-value. In Q learning and related algorithms, an agent tries to learn the optimal policy from its history of interaction with the environment. The learning rate α determines to what extent the newly acquired information will override the old information. A factor of 0 will make the agent not to learn anything, while a factor of 1 would make the agent consider only the most recent information. The discount factor γ determines the importance of future rewards. A factor of 0 will make the agent "opportunistic" by only considering current rewards, while a factor approaching 1 will make it strive for a long-term high reward.

Q learning algorithm:

Controller Q – learning (S, A, γ, α)

Inputs

S is a set of states
A is a set of actions
γ is the discount
α is the learning rate

Local

Real array Q[s, a]
Previous state s
Previous action a

Initialize Q[S, A] arbitrarily
Observe current state s

Repeat
Select and carry out an action a
Observe reward r and state s'
$Q[s,a] \leftarrow Q[s,a] + \alpha(r +$
$\gamma \max_{a'} Q[s',a'] - Q[s,a])$
$s \leftarrow s'$ until termination

First pole angle and angular velocities have been divided from -0.209 to 0.209 Radian and -2.0933 to 2.0933 Radian/sec. Second pole angle and angular velocities have been divided from -0.1046 to 0.1046 Radian and -1.046 to 1.046 Radian/sec. The state space for double inverted pendulum in Q learning has been shown in Table 1. Pole angles θ_1 and θ_2 are taken in Radian. Pole angular velocities $\dot{\theta}_1$ and $\dot{\theta}_2$ are taken in Radian/second.

Table 1. State space for double inverted pendulum in Q learning

θ_1	<-0.21	-0.21, -0.11	-0.11, 0	0, 0.11	0.11, 0.21	>0.21
$\dot{\theta}_1$	<-2.09		-2.09, 2.09		>2.09	
θ_2	<-0.104		-0.104,0.104		>0.104	
$\dot{\theta}_2$	<-1.05		-1.05, 1.05		>1.05	

Double inverted pendulum system has been divided into 6*3*3*3*3= 162 states and action size is just two: clockwise torque (τ_1) +10 Nm and anticlockwise torque -10 Nm. The sampling interval is 0.02 second. The length of first and second rigidless link is 0.2 m. The first pendulum is kept within ± 0.209 Radian and second pendulum is kept within ± 0.1046 Radian. The stiffness and damping coefficient of joint 1 is 0.01 N/Radian and 0.001 NSec/Radian.

The change of mass m_2 from Double support phase (DSP) to Single Support Phase (SSP) is a function of θ_1 and θ_2.

$m_{1L} = m_{1R} = 0.5$ Kg
$m_{2L} = m_{2R} = 4$ Kg
$m_2 = m_{2L} + m_{2R} = 8$Kg
Δm_2 = Change in mass of m_2 from SSP to DSP
= 8.5-4
= 4.5 Kg

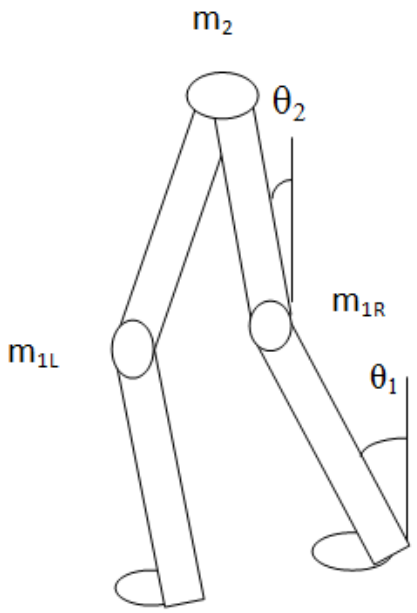

Fig. 3 Bipedal Robot

Figure 3 shows the model of bipedal robot. There is an increase of θ_1 and θ_2 when the bipedal robot is changing from double support phase to single support phase.

IV. SIMULATION

The double inverted pendulum task was simulated for 14,000 iterations using Q learning in Reinforcement Learning and it was tested for various parameters: learning rate α and discount factor γ. The learning rate parameter α and discount factor γ are important factors for the agent to learn. Various combinations of learning rate and discount factor (0.4, 0.8), (0.5, 0.85) and (0.6, 0.9) was tested. A combination of learning rate and discount factor of 0.5 and 0.85 resulted the agent to control the double inverted pendulum successfully. Simulation of double inverted pendulum using Q learning has been done.

Limit cycle is the study of dynamical systems with two dimensional phase space. Limit cycles occur in non-linear systems. Finally, pole angle and pole angular velocity become zero. Therefore, the double inverted pendulum becomes stable using Q learning in Reinforcement Learning. The results of simulation of double inverted pendulum using Q learning has been shown in Figure 4. This shows the variation of $\dot{\theta}_1$ vs θ_1 in Q learning of double inverted pendulum in Q learning.

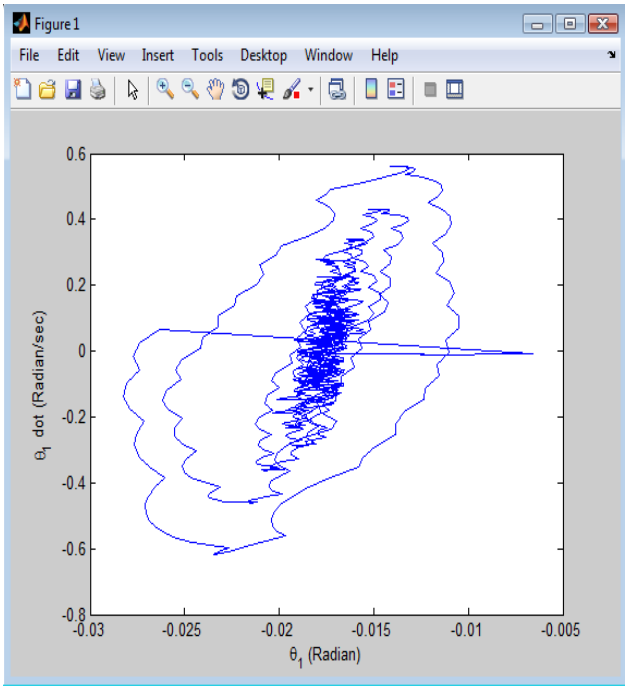

Fig.4 Plot of $\dot{\theta}_1$ vs θ_1 in Q learning

The variation of $\dot{\theta}_2$ vs θ_2 in Q learning of double inverted pendulum has been shown in Figure 5. Finally, pole angle and pole angular velocity become zero. Therefore, the double inverted pendulum becomes stable using Q learning in reinforcement Learning.

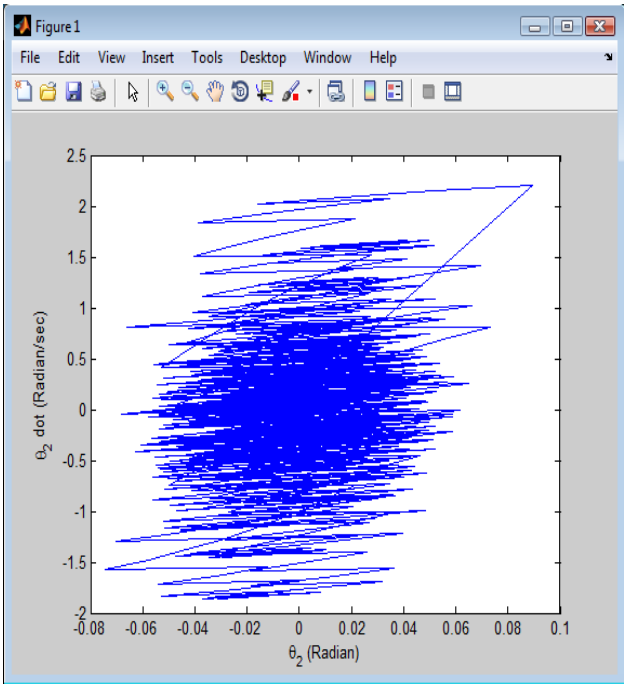

Fig. 5 Plot of $\dot{\theta}_2$ vs θ_2 in Q learning

V. CONCLUSIONS

In Q learning, an agent tries to learn an optimal policy from its history of interaction with the environment. A history of an agent is a sequence of state-action rewards. Limit cycle behaviour of double inverted pendulum has been shown in this paper. The objective of this work is to show how double inverted pendulum can be balanced using Q learning in reinforcement learning. Finally, pole angles and angular velocities become zero. This indicates that the double inverted pendulum is stable using Q learning in Reinforcement learning.

REFERENCES

[1] Yibiao Zhao, Siwei Luo, Liang Wang, Aidong Ma and Rui Fang, “Vague neural network based reinforcement learning control system for inverted pendulum”, ICONIP 2006, part III, LNCS 4234, pp.692-701, 2006.

[2] Alexander Bogdanov, “Optimal control of a double inverted pendulum on a cart”, Technical report CSE-04-006, December 2004.

[3] Dymitry Kaynov, “Open motion control architecture for humanoid robots”, Doctoral thesis, Universidad carlos III de Madrid, 2008.

[4] Yuan shaoqiang, Liu Zhong, Li Xingshan, “Modeling and Simulation of robot based on Matlab/Sim Mechanics”, Proceedings of the 27 th Chinese Control Conference July 16-18, 2008, Kumming, Yunnan, China.

[5] Kui-Hong Park, Jun Jo and Jong-Hwan Kim, “Stabilization of Biped Robot based on two mode Q learning”, 2 nd International Conference on Autonomous Robots and Agents, December 13-15, 2004, Palmerston North, New Zeland.

[6] Wal Suleiman, Fumio Kanehiro, Kanako Miura and Elichi Yoshida, “Enhancing Zero moment Point Control: System Identification Approach”, Advanced Robotics Volume 25, November 3-4, 2011, pp. 427-446.

[7] Dong W. Kim, Nak-Hyun Kim, Gwi Tae Park, “ZMP based neural network inspired humanoid robot control”, in Nonlinear Dynamics, Volume 67, April, 2011, pp.793-806.

[8] Fumihiko Asano, Masaki Yamakita, Norihiro Kamamichi, and Zhi-Wei Luo, “A novel gait generation for biped walking robots based on mechanical energy constraint”, in IEEE transcations on robotics and automation, Vol. 20., no-3, June 2004.

[9] Yonghwan Oh. Kyung-ho Ahn, Doikkimand changhwan Kim, “An analytical method to generate walking pattern of humanoid robot”, IEEE industrial electronics, IECON 2006, pp. 4159-4164, 2006.

[10] Zhe Tang and MengJOOEr, “Humanoid 3D Gait generation Based on Inverted Pendulum Model”, 22nd IEEE International Symposium on Intelligent control, Singapore, 1-3 October, 2007.

[11] A. A. Thant and K. K. Aye, “Application of Cubic Spline Interpolation to Walking Patterns of Biped Robot”, World Academy of Science, Engineering and Technology, 2009.

[12] V. Gullapalli, J A Franklin, H. Benbrahim, “Aquiring Robot Skill via Reinforcement Learning”, IEEE Control System Magazine, February, 1994.
[13] F. F. Stulp, J. Buchli, E. Theodorou, s. Schaal, “Reinforcement Learning of Full-body Humanoid Motor Skills”, 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 405-410, 2010.
[14] A. Enoch, A. Sutas, S. Nakaoka, S. Vijaykumar, “BLUE: A Bipedal Robot with Variable Stiffness and Damping”, 12th IEEE-RAS International Conference on Humanoid Robots, Osaka, Japan, 2012.
[15] J. Nakanishi, J. Morimoto, G.endo, G. Cheng, S. Schaal and M. Kawato, “A framework for learning biped locomotion with dynamical movement primitives”, IEEE-RAS International Conference on Humanoid Robots, Los Angles, USA, 2004.
[16] H. Benbrahim, J. A. Franklin, “Biped dynamic walking using Reinforcement learning”, Robotics and Autonomous Systems, Vol. 22, pp. 283-302, 1997.