

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/381822083>

Q-Learning Based Control for Swing-Up and Balancing of Inverted Pendulum

Conference Paper · May 2024

DOI: 10.1109/IETC61393.2024.10564347

CITATIONS

0

READS

36

3 authors, including:



Antora Dev

Idaho State University

7 PUBLICATIONS 98 CITATIONS

[SEE PROFILE](#)



Marco P. Schoen

Idaho State University

154 PUBLICATIONS 1,415 CITATIONS

[SEE PROFILE](#)

Q-learning based Control for Swing-up and Balancing of Inverted Pendulum

Antora Dev
Department of Electrical
and Computer Engineering
Idaho State University
Pocatello, ID, USA
antoradev@isu.edu
0000-0002-1968-9368

Kanan Roy Chowdhury
Measurement & Control Engineering
Research Center (MCERC)
Idaho State University
Pocatello, ID, USA
kananroychowdhury@isu.edu
0000-0002-9481-0244

Marco P. Schoen
Measurement & Control Engineering
Research Center (MCERC)
Idaho State University
Pocatello, ID, USA
schomarc@isu.edu
0000-0002-6572-0119

Abstract—This study addresses the classic control problem of stabilizing an inverted pendulum on a moving cart, a challenge in control theory and robotics due to its inherent instability and highly nonlinear dynamics. We explore the application of Q-learning, a model-free reinforcement learning algorithm, and its efficacy in deriving an optimal control policy for the system without precise system models. Our approach utilizes Q-learning's capacity for stabilizing a pendulum in an upright position on the top of the horizontally moving cart within a certain boundary. Our strategy adapts to a dynamic environment while showcasing its robustness in developing control policies for complex systems. This research bridges classical control theory with reinforcement learning techniques, contributing to the domain by demonstrating the versatility and potential of machine learning in control tasks.

Keywords—Inverted pendulum, Q-learning, Reinforcement learning, non-linear dynamics, control system.

I. INTRODUCTION

The problem of swinging-up and balancing an inverted pendulum on a moving cart is a unique challenge in control theory and robotics. This system, characterized by its inherently unstable and highly nonlinear dynamics, serves as a fundamental benchmark for testing control strategies and algorithms [1]. The complexity of the inverted pendulum problem lies in maintaining the pendulum in an upright position while managing the movement of the cart horizontally.

In recent years, the field of machine learning, particularly reinforcement learning, has shown promising results in solving various complex control problems [2]. Reinforcement learning (RL), a subset of machine learning, is particularly relevant. RL algorithms learn to make decisions by interacting with the environment, making them suitable for control tasks where precise models of the system dynamics are not available.

Among various RL algorithms, Q-learning has emerged as a robust method for learning control policies. It is a value-based algorithm that learns the value of taking certain actions in specific states from surroundings, thereby learning an optimal policy indirectly. Q-learning's ability to handle problems with continuous state spaces and its model-free nature make it an excellent candidate for the inverted pendulum problem [3].

This paper explores the application of Q-learning in a model-free environment for an inverted pendulum which is particularly applied to unknown dynamics and is difficult to handle precisely. In our study, we implemented a unique learning algorithm that learns an optimal control policy for

swinging-up and balancing that pendulum on the top of a horizontally moving cart within a limited boundary.

The structure of the remainder of this paper is organized to facilitate a comprehensive understanding and discussion of the study. Section II explores existing literature, setting the stage for our investigation. Section III lays the foundational concepts of inverted pendulum mechanics and the mechanics of Q-learning as a control solution. In Section IV, we present the methodology employed in our approach. Section V presents the outcomes of our analysis and engages in a thorough discussion of the results. Finally, the paper concludes with Section VI, which summarizes the findings and proposes directions for future research.

II. RELATED RESEARCH

Previously, Watkins and Dayan *et al.* [4] introduced Q-learning, a form of model-free reinforcement learning by addressing "learning from delayed reward". Further studies have explored modifications to the basic Q-learning algorithm to enhance performance on the inverted pendulum task. The action critic in the model-free algorithm was later proposed by Lillicrap *et al.* [5] with Deterministic Policy Gradients (DPG) and represented a significant advancement over continuous action spaces.

Ahmad *et al.* [6] did a comparative analysis between conventional control system (i.e. root-locus-based control, state compensator control, and proportional-derivative (PD) control) to a reinforcement learning strategy, particularly the proximal policy optimization (PPO) and found that with sufficient training, RL can match the performance of traditional controls without prior knowledge of the system's dynamics, making it suitable for the environment with little model information available. On the other hand, Khoa *et al.* [7] presented a modified deep Q-network (DQN) algorithm for controlling an inverted pendulum that allowed a range of force outputs to improve the system performance. Poorna *et al.* [8] developed a model environment and trained neural network controllers using reinforcement learning algorithms to balance a pendulum on a vertically moving cart within displacement limits.

In addition, Wang *et al.* [9] investigated reinforcement learning control algorithms, demonstrating the effective application of Q-learning optimized by a BP neural network for real-time control of a one-stage inverted pendulum, which

showcased improved stability and effectiveness. Zeynivand *et al.* [10] integrated Q-learning with Proportional-Integral-Derivative (PID) control to manage the dynamics of a double inverted pendulum, highlighting the enhanced performance through simulation results. Safeea *et al.* [3] evaluated a Q-learning approach to the continuous control problem of robot inverted pendulum balancing, emphasizing the use of discrete action space reinforcement learning to handle continuous control tasks efficiently, further showing the method's applicability in real-world robotic systems through simulation to real transfer techniques.

Our approach demonstrates how RL in a model-free environment can be effectively applied to a continuous state-space problem that emphasizes the importance of state-space discretization and its impact on learning efficiency and control precision. We have also analyzed the trade-off between the learning rate, the discount factor, and the optimized control strategy in terms of total cumulative rewards, the pendulum staying upright position, and the maximum displacement of the cart within the boundaries.

III. THEORETICAL BACKGROUND

A. Inverted Pendulum Dynamics

The system dynamics of an inverted pendulum on a cart are governed by the relation of motion, which describes how the movement of the pendulum and the cart can change over time due to forces and torques applied. The dynamics are typically modeled by a set of differential equations derived from Newton's second law of motion.

For the cart, the horizontal force applied to the cart translates into acceleration, considering friction and the reaction force exerted by the pendulum's weight. The pendulum experiences torque due to gravity, which tends to return it to the downward-hanging equilibrium. The motion of the cart can add or subtract from the pendulum's angular acceleration depending on the direction of the cart's movement and the pendulum's position.

In our system, we have discretized four states i.e. cart position, cart velocity, pendulum angle, and pendulum angular velocity. The control input, in the form of force applied to the cart, influences the next state of the system based on these dynamics.

The overall goal is to apply control inputs that bring the pendulum into an upright position and then balance it on the top of the cart despite the inherently unstable dynamics of the system. The learning algorithm adjusts its strategy based on the rewards received for keeping the pendulum upright, optimizing the force inputs over time to achieve this balance.

B. Q-learning Algorithm

Q-learning is a model-free reinforcement learning algorithm that learns a policy, which an agent should take under a given state by their action to maximize a cumulative reward. It doesn't require knowing a model of the environment and can handle problems with stochastic transitions and rewards, without needing adaptations. In Q-learning, the agent learns a function $Q(s, a)$, which represents the expected utility of taking action a in state s [11]. This function is updated iteratively using the Bellman equation:

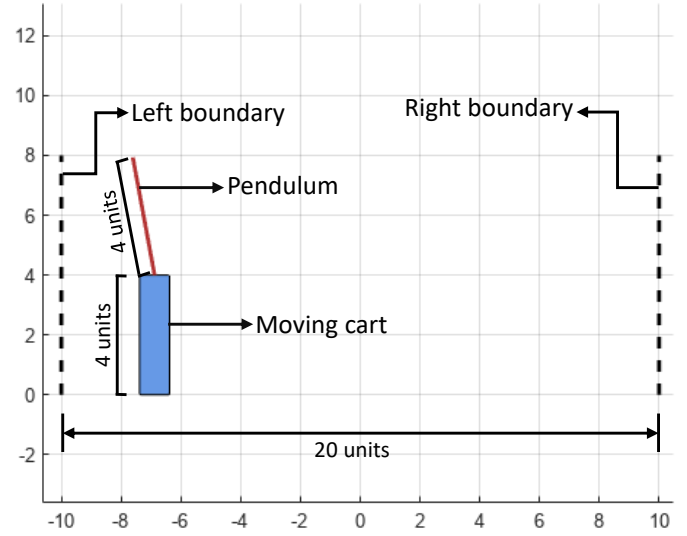


Fig. 1. The inverted pendulum on a moving cart system.

Algorithm 1: Pseudo-code for Inverted Pendulum on a Moving Cart in Reinforcement Learning

Input: g (gravity), l (pendulum length), dt (time step), ep (number of episodes), α (learning rate), γ (discount factor), f_s (sensitivity to force), f_{max} (maximum applicable force), s_t (initial state), θ_l (angle limit).

Output: Q (maximum Q values), r_t (rewards per episode), $maxpos$ (max positions per episode), $duration$ (time within angle limit per episode).

```

1  $Q \leftarrow$  Initialize Q-table with zeros;
2  $policy \leftarrow$  set epsilon parameters;
3 for episode  $\leftarrow 1$  to  $ep$  do
4    $state \leftarrow x, v, \theta, \omega$ ;
5    $r_t, maxpos(+), maxpos(-) \leftarrow 0$ ;
6    $s_t \leftarrow$  Observe the initial state;
7   for  $t \leftarrow 1$  to time do
8     select  $a_t$  based on current policy;
9      $s_{t+1} \leftarrow$  apply  $a_t$  with  $f_s$  and  $f_{max}$ ;
10     $r_t \leftarrow$  get reward based on  $s_{t+1}$ ;
11     $Q \leftarrow$  update  $Q$  for  $r_t$  and best  $Q(s_{t+1})$ ;
12     $r_t \leftarrow r_t +$  reward obtained at this  $dt$ ;
13     $maxpos \leftarrow$  update  $maxpos$  based on  $s_{t+1}$ ;
14    time  $\leftarrow$  update time spent within  $\theta_l$ ;
15  end
16   $r_t[episode] \leftarrow$ 
17    total reward observed during episode;
18   $maxpos[episode] \leftarrow$ 
19    maximum of  $maxpos(+)$  and  $maxpos(-)$ ;
20   $duration[episode] \leftarrow$ 
21    cumulative time spent within  $\theta_l$ ;
22 end

```

$$Q_{new}(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (1)$$

Where:

- α is the learning rate.
- r_{t+1} is the reward received after transitioning from state s_t to s_{t+1} .

- γ is the discount factor, which balances immediate and future rewards.
- $\max Q(s_{t+1}, a)$ is the estimated optimal future value.

IV. PROPOSED STRATEGY

A. Model Simulation Mechanism

Our simulation framework for the inverted pendulum on a moving cart was predicated on the accurate representation of the physical dynamics governing the system, combined with a reinforcement learning algorithm to control the cart's movements. For the movement of the cart dynamics, the cart's acceleration is derived from Newton's second law and takes into account the force applied, the gravitational force acting on the pendulum, and the pendulum's inertia. The visualization of the whole system dynamics is shown in figure 1. The force applied to the cart is proportional to the deviation of the pendulum's angle from the vertical, scaled by the force sensitivity parameter. The possible actions are:

$$\text{force} = \begin{cases} -\text{force_sensitivity}, & \text{if action} = 1 \\ 0, & \text{if action} = 2 \\ \text{force_sensitivity}, & \text{if action} = 3 \end{cases}$$

where:

- action 1: Apply a leftward force
- action 2: Apply no force (0)
- action 3: Apply a rightward force

We have also set a boundary on each side with an equal distance from the initial position of the cart, which plays a crucial role in the pendulum's stabilization process. This interaction induces a directional change in the cart, providing a momentum shift that aids in maintaining the pendulum's upright position within 20 degrees. Figure 2 illustrates the impact of the system dynamics before hitting the left-sided boundary barrier, poised to reverse direction due to the applied backward force. For the pendulum, angular acceleration is computed using the Euler-Lagrange equation which considers the effects of gravity, the reaction force from the cart, and the pendulum's angular velocity. The pendulum's horizontal acceleration occurs due to the cart's motion, and the rotational dynamics of the pendulum are influenced by the cart's acceleration and the gravitational pull on the pendulum.

B. State Initialization

The subsequent state of the inverted pendulum system is determined by the current state and the taken action. The updated parameters are derived from the following simplified physical equations:

$$\dot{x} = x + dt \cdot v \quad (2)$$

$$\dot{v} = v + dt \cdot a \quad (3)$$

$$\dot{\theta} = \theta + dt \cdot \theta v \quad (4)$$

$$\dot{\omega} = \omega + dt \cdot \alpha_m \quad (5)$$

where:

- x is the initial position of the cart,

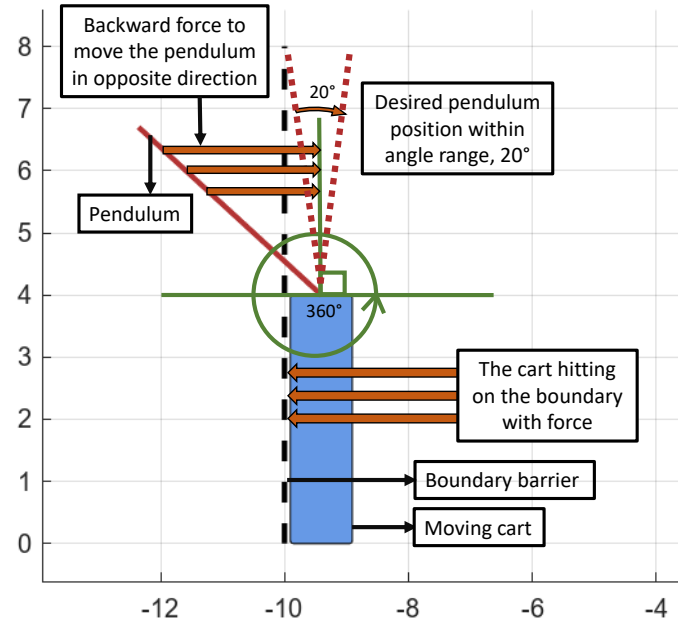


Fig. 2. Boundary-induced stabilization of the inverted pendulum within 20 degrees on a moving cart before hitting the left-side boundary barrier.

- \dot{x} is the updated position of the cart,
- v is the initial velocity of the cart,
- \dot{v} is the updated velocity of the cart,
- θ is the initial angle of the pendulum from the vertical,
- $\dot{\theta}$ is the updated angle of the pendulum from the vertical,
- ω is the initial angular velocity of the pendulum,
- $\dot{\omega}$ is the angular velocity of the pendulum,
- a is the linear acceleration of the cart, and
- α_m is the angular acceleration of the pendulum.

The acceleration and the angular acceleration are calculated using the force exerted and the dynamics of the pendulum, which include gravitational force and the pendulum's inertia.

C. Implementation of Q-learning

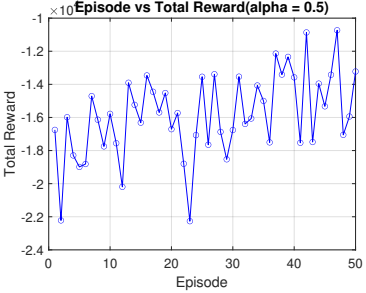
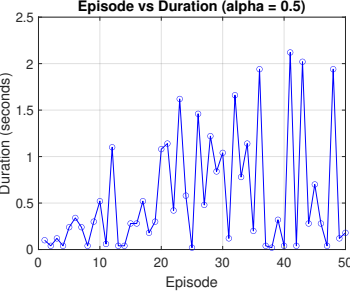
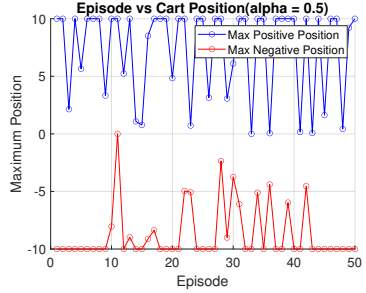
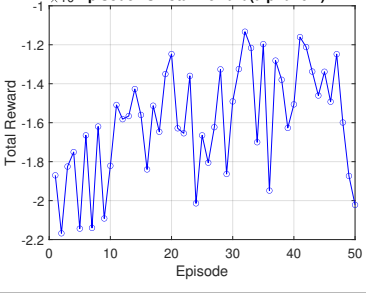
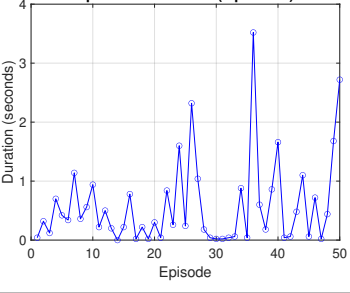
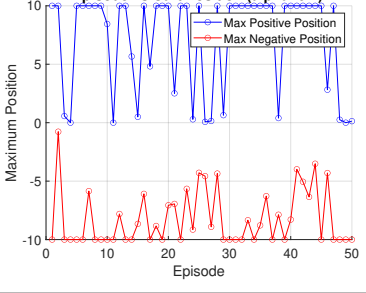
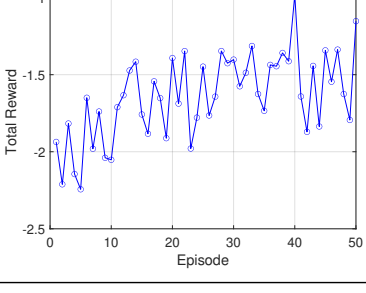
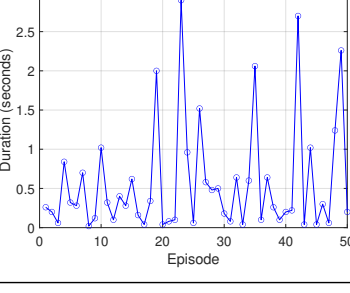
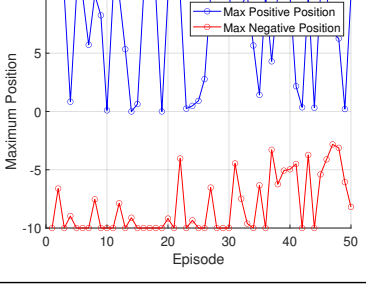
In our research, we implemented a Q-learning model to control an inverted pendulum on a moving cart within a simulation environment. The model was initialized by setting gravitational acceleration g to 9.81 ms^{-2} , the length of the pendulum, l to 4.0m , and the simulation time step, dt to 0.02s . The learning process was configured to run over 500 episodes, with a learning rate α and a discount factor γ . We analyzed our system with three different learning rates (0.5, 0.7, and 0.9) and with three different discount factors (0.1, 0.2, and 0.5) to find the balanced parameters for the model in terms of total cumulative rewards, staying the pendulum in the upright position, and the maximum cart displacement.

The state-space, consisting of the cart's position and velocity as well as the pendulum's angle and angular velocity, was discretized into 50 bins each, allowing the Q-learning algorithm to operate on a finite representation of the continuous state-space. The exploration rate ϵ was set to decay from 0.9 to 0.01 across 80% of the episodes, encouraging the model to explore actions initially and gradually shift to exploiting the learned policy.

The core of the implementation was the Q-learning loop, where for each episode, the system was initialized with the

TABLE I

COMPARATIVE ANALYSIS OF THE REWARD FUNCTION, MAXIMUM DURATION OF STAYING WITHIN -10 TO 10 DEGREES, AND MAXIMUM CART DISPLACEMENT PER EPISODE FOR INVERTED PENDULUM ON A MOVING CART SYSTEM IN TERMS OF CHANGES IN LEARNING RATE.

Episode Vs Parameters of Inverted Pendulum System			
	Total Cumulative Reward	Staying Upright Duration	Maximum Cart Displacement
Changes of Learning rate, α	<p>$\alpha = 0.5$</p> 		
	<p>$\alpha = 0.7$</p> 		
	<p>$\alpha = 0.9$</p> 		

pendulum slightly deviated from the upright position. Actions were chosen using an epsilon-greedy policy, allowing for both explorative and exploitative decisions. The force applied to the cart was either -500 , 0 , or $+500$, corresponding to the leftward force, no force, and rightward force respectively, as determined by the policy. These actions were sensitive to the angular deviation of the pendulum, aiming to correct its position.

The next states of the system were calculated by incorporating the dynamics of the pendulum and the applied forces. If the cart reached the boundaries of ± 10 units on either the left side or the right side from the initial position, it incurred a penalty to the reward function and provided the backward forces of 500 and -500 in the left and right side boundaries respectively. Rewards were used to update the Q-values in the Q-table, reflecting the learned policy's efficacy.

Throughout the simulation, data was recorded on the total rewards, the maximum cart displaces, and the duration of the pendulum remained within the desired angular limit in the upright position. This data visualization and analysis is discussed at the end of the simulation, providing insight into the learning progression and the control strategy's success.

V. RESULTS AND DISCUSSION

A. Impact of Learning Rate

In this section, we emphasized the significant role of the learning rate in the Q-learning algorithm's efficacy. Table I displays the system's performance with varying α values of 0.5 , 0.7 , and 0.9 . Observing the total cumulative reward, it is evident that a learning rate of $\alpha = 0.7$ results in a more stable reward convergence, with less variance in reward values across episodes even though the highest reward seems to be achieved by $\alpha = 0.9$. The maximum staying upright between -10 and 10 degrees duration, indicating the pendulum's stability, is achieved at $\alpha = 0.7$ (around 3.6 seconds), implying sustained equilibrium for longer periods. The maximum cart displacement within the boundaries shows less extreme values with $\alpha = 0.7$, suggesting a more nuanced and effective control approach within the limits.

B. Impact of Discount Factor

In this section, we analyzed the pivotal influence of the discount factor on the stability and learning efficacy of an inverted pendulum, shown in Table II. The performance of the system is evaluated at varying levels of γ - 0.1 , 0.2 , and 0.5 . Upon examining the Total Cumulative Reward, we discern

TABLE II

COMPARATIVE ANALYSIS OF THE REWARD FUNCTION, MAXIMUM DURATION OF STAYING WITHIN -10 TO 10 DEGREES, AND MAXIMUM CART DISPLACEMENT PER EPISODE FOR INVERTED PENDULUM ON A MOVING CART SYSTEM IN TERMS OF CHANGES IN DISCOUNT FACTOR.

Episode Vs Parameters of Inverted Pendulum System			
Changes of discount factor, gamma			
	Total Cumulative Reward	Staying Upright Duration	Maximum Cart Displacement
	<p>gamma = 0.1</p>	<p>Episode vs Duration(gamma=0.1)</p>	<p>Episode vs Cart Position(gamma=0.1)</p>
	<p>gamma = 0.2</p>	<p>Episode vs Duration(gamma=0.2)</p>	<p>Episode vs Cart Position (gamma=0.2)</p>
	<p>gamma = 0.5</p>	<p>Episode vs Duration(gamma=0.5)</p>	<p>Episode vs Maximum Cart Position(gamma=0.5)</p>

that $\gamma=0.2$ facilitates an enhanced reward accumulation over episodes. Correspondingly, the highest duration in which the pendulum remains in the upright position within -10 to 10 degrees is marked at $\gamma=0.2$. This suggests that the system maintains its balance for a much longer time when the discount factor is set to 0.2. Similarly, the maximum cart displacement, a reflection of the control strategy's effectiveness at $\gamma=0.2$, indicates a more refined and calibrated response for its control mechanism. Therefore, a discount factor of 0.2 is identified as an optimal and delicate balance between immediate and future rewards for the inverted pendulum system.

C. Duration of Stability

According to our above analysis, we have set our best choice of parameters for the learning rate (0.7) and discount factors (0.2). Figure 3 shows our observation on the duration of the stability of the inverted pendulum between ± 10 -degree angle from the vertical in each episode over time which indicates the system's stability over time. Initially, the pendulum maintains its stability within the angle limit for short periods, reflecting the learning phase of the Q-learning algorithm. As the episodes progress, we have found that the pendulum remains stable for a longer time, suggesting

improvement in the learning strategy. The highest duration for staying upright position is found around 3.25 seconds. However, the variability in stability duration also indicates that the learning process experiences fluctuations, which could be due to exploration or the complexity of the system's dynamics.

D. Reward Optimization

Figure 4 refers to the total reward accumulated per episode. A higher total reward is indicative of better performance, i.e. the system is more frequently achieving the desired state. From there it is noticeable that the reward trend shows slow but a clear upward trajectory, which implies that the algorithm is gradually improving its reward function.

E. Analysis of Cart Displacement

In this section, we have analyzed the maximum cart movement for both positive and negative directions in each episode. Ideally, a successful control strategy would minimize these extremes, keeping the cart near the center. The graph in Figure 5 indicates that before around the 150th episode, the cart frequently hit the boundaries, as denoted by the spikes to the extremes of the position range; however, after 150 episodes, the cart appeared to be hitting only one of the sides. This

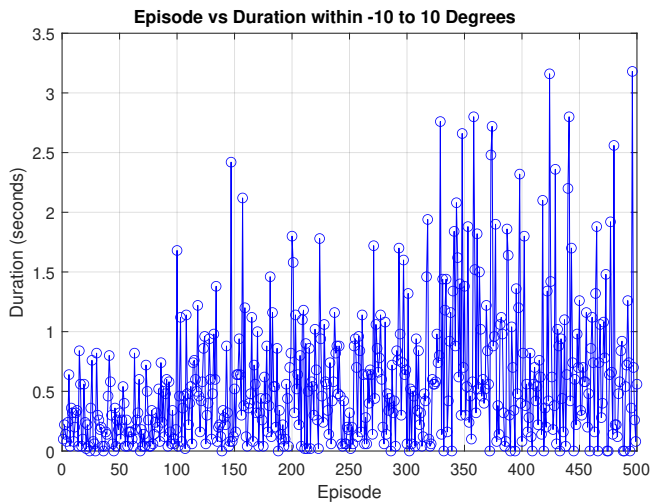


Fig. 3. Episode Vs the duration of staying the pendulum in the upright position with -10 to 10 degrees.

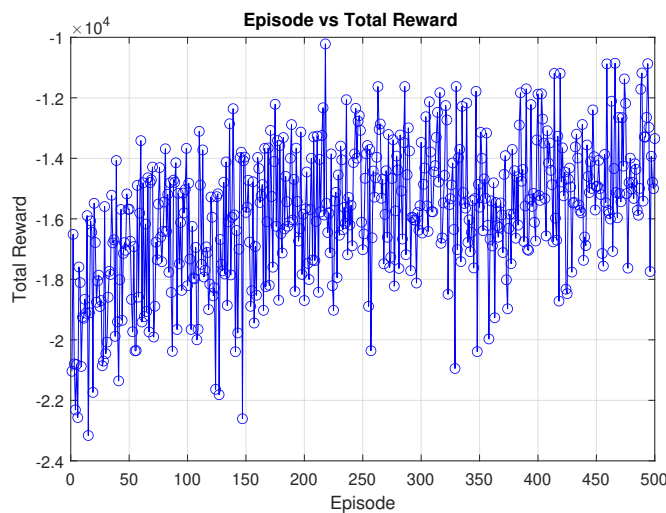


Fig. 4. Episode Vs Total Cumulative Reward during learning.

behavior could be due to the given maximum force to the boundaries as an exploration strategy to keep the pendulum in the upright position for a longer time which suggests further tuning of the learning parameters or additional episodes could be beneficial.

VI. CONCLUSION AND FUTURE WORK

This research has successfully demonstrated the use of Q-learning for swinging up and balancing an inverted pendulum on a moving cart. Through a series of experiments, we have determined the optimal parameters that yield the best performance in terms of stability and learning efficiency. Specifically, the optimal discount factor was identified to significantly influence the learning outcome and the system's stability. For future work, we plan to explore the integration of advanced reinforcement learning techniques, such as Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO), which have the potential to handle the complexities of the control problem more effectively. Additionally, we aim to implement an adaptive learning rate mechanism that could potentially accelerate the learning process. This work not only

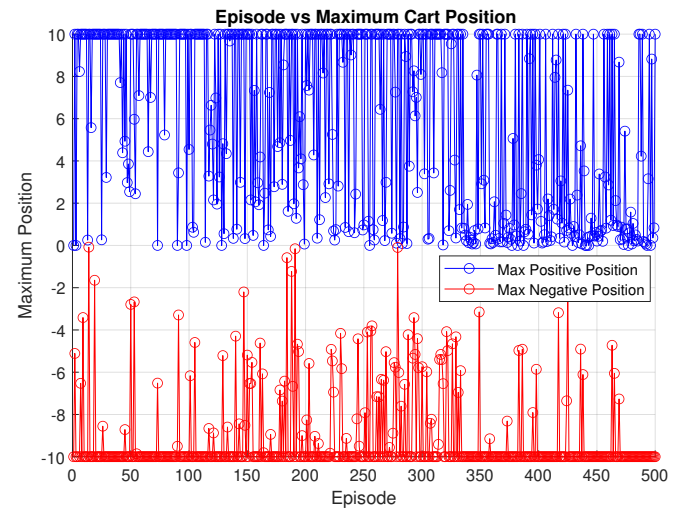


Fig. 5. Episode Vs the maximum movement of cart position within -10 to 10 units of boundaries.

advances the understanding of Q-learning in control systems but also stands as a testament to the potential of reinforcement learning in the field.

ACKNOWLEDGEMENT

This work was supported through a Center of Advanced Energy Studies (CAES) grant. The support is greatly appreciated.

REFERENCES

- [1] E. Sivaraman and S. Arulselvi, "Modeling of an inverted pendulum based on fuzzy clustering techniques," *Expert Systems with Applications*, vol. 38, pp. 13 942–13 949, 2011.
- [2] G. G. Jaman, A. Monson, K. R. Chowdhury, M. Schoen, and T. Walters, "System identification and machine learning model construction for reinforcement learning control strategies applied to lens system," pp. 1–6, 2022.
- [3] M. Safeea and P. Neto, "A q-learning approach to the continuous control problem of robot inverted pendulum balancing," *Intelligent Systems with Applications*, vol. 21, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2667305323001382>
- [4] C. J. C. H. Watkins and P. Dayan, "Technical note q-learning," vol. 8, pp. 279–292, 1992.
- [5] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: <http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#LillicrapHPHETS15>
- [6] A. Ataka, A. Sandiwan, H. Thunay, D. R. Utomo, and A. I. Cahyadi, "Inverted pendulum control: A comparative study from conventional control to reinforcement learning," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 12, pp. 197–204, 2023.
- [7] K. N. Dang, V. T. Thi, and L. V. Van, "Development of deep reinforcement learning for inverted pendulum," *International Journal of Electrical and Computer Engineering*, vol. 13, pp. 3895–3902, 2023.
- [8] P. H. Vamsi A, M. D. Ratolilar, and R. P. Kumar, "Swinging up and balancing a pendulum on a vertically moving cart using reinforcement learning," pp. 1668–1673, 2021.
- [9] L. Wang, Y. Liu, and X. Zhai, "Design of reinforce learning control algorithm and verified in inverted pendulum," *Chinese Control Conference, CCC*, pp. 3164–3168, 2015.
- [10] A. Zeynivand and H. Moodi, "Swing-up control of a double inverted pendulum by combination of Q-Learning and pid algorithms," *2022 8th International Conference on Control, Instrumentation and Automation, ICCIA 2022*, 2022.
- [11] "Reinforcement learning 101: Q-learning | towards data science," <https://towardsdatascience.com/reinforcement-learning-101-q-learning>.