Reinforcement Learning Algorithms Applied to Grid World and Simple Pendulum

Raghavendra S Navaratna
rsn3@illinois.edu
Link to the repository:

I. Introduction

This is a report on evaluating the model-based and model-free reinforcement algorithms applied to the grid world problem and a simple pendulum.

II. Grid World

A. Policy Iteration Method

Policy iteration is a model-based reinforcement learning algorithm that iteratively improves the policy and value function until convergence by alternating between policy evaluation and policy improvement. [1]

TABLE I
Hyperparameter values

Hyperparameter	Values
Gamma (γ)	0.95
Epsilon (ϵ)	0.1
Alpha (α)	0.5
Iterations	100
Threshold (θ)	1e-8

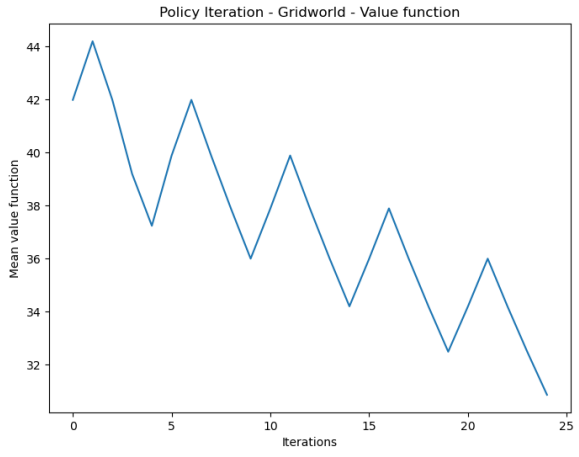

Fig. 1. Mean Value Function vs. Number of Iterations

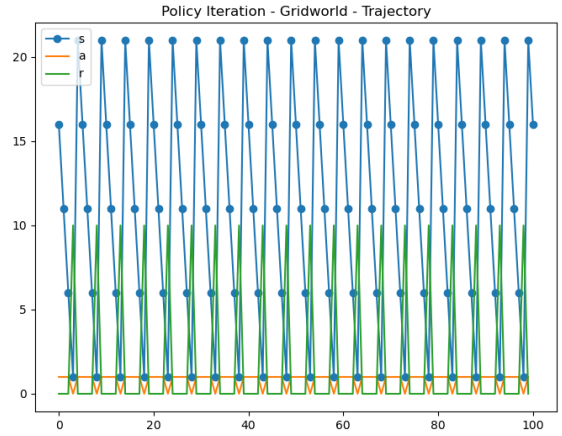

Fig. 2. Policy vs. Trajectory

B. Value Iteration Method

Value iteration is a dynamic programming algorithm that computes the optimal value function by iteratively updating state values based on the maximum expected future reward achievable from each state. It converges to the optimal policy after a finite number of iterations. [1]

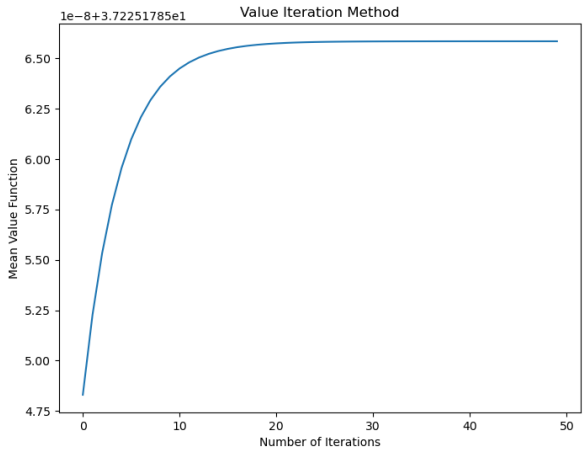

Fig. 3. Mean Value Function vs. Number of Iterations

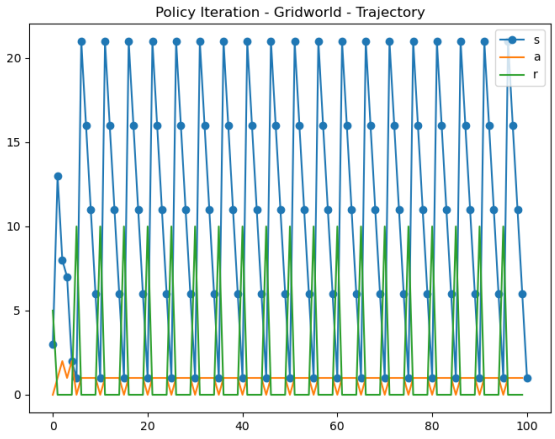

Fig. 4. Policy vs. Trajectory

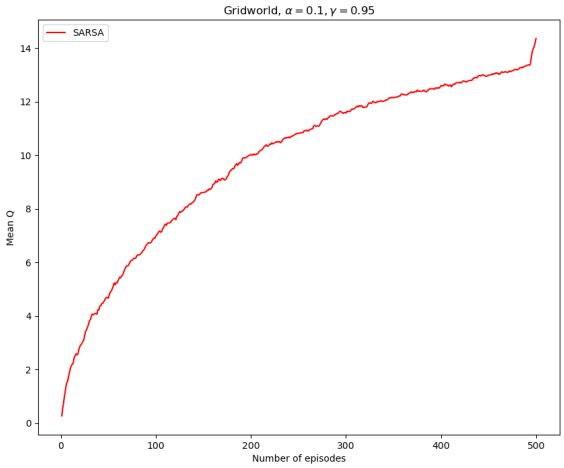

Fig. 5. Learning Curve - Mean Value Function vs. Number of Episodes

C. SARSA

SARSA (State-Action-Reward-State-Action) is a model-free reinforcement learning algorithm used for estimating the Q-values of a policy. It updates the Q-value of the current state-action pair based on the reward obtained and the Q-value of the next state-action pair, using the current policy to select the next action. [1]

TABLE II
HYPERPARAMETER VALUES

Hyperparameter	**Values**
Gamma (γ)	0.95
Epsilon (ϵ)	0 - 1
Alpha (α)	0.2 - 1
Iterations	100
Episodes	500
Threshold (θ)	1e-8

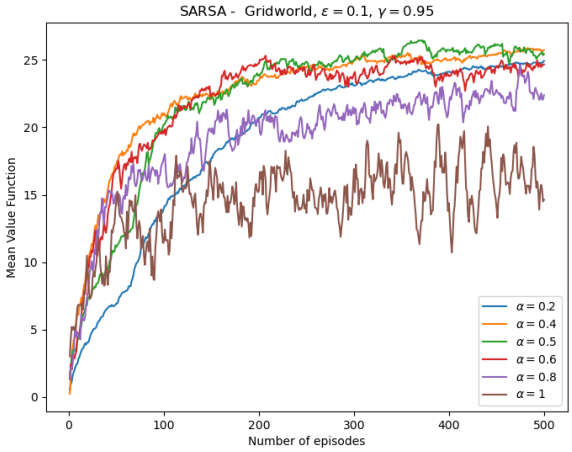

Fig. 6. Learning Curve for Various Learning Rate (α)

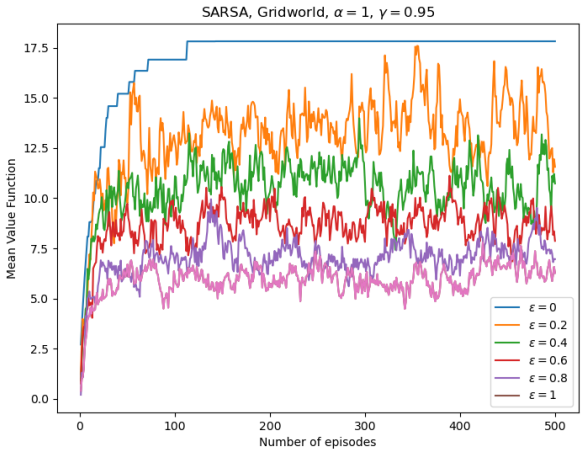

Fig. 7. Learning Curve for Various (ϵ)

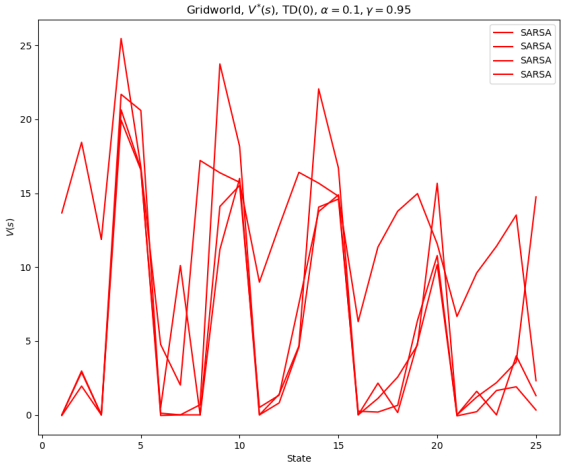

Fig. 9. State-Value Function vs. Number of Episodes

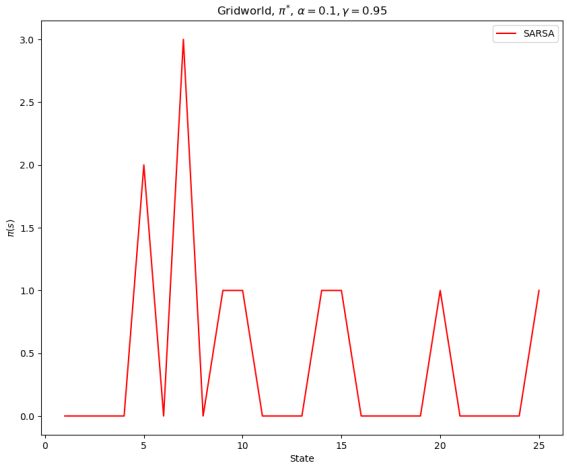

Fig. 8. Action-Value Function vs. Number of Episodes

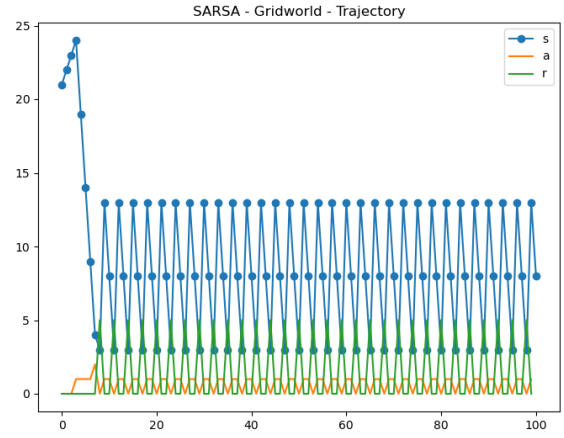

Fig. 10. Policy vs. Trajectory

D. Q-Learning

Q-learning is a model-free reinforcement learning algorithm that learns an optimal policy without knowledge of the environment's dynamics. It estimates the optimal action-value function by iteratively updating Q-values based on experience gained through exploration. [1]

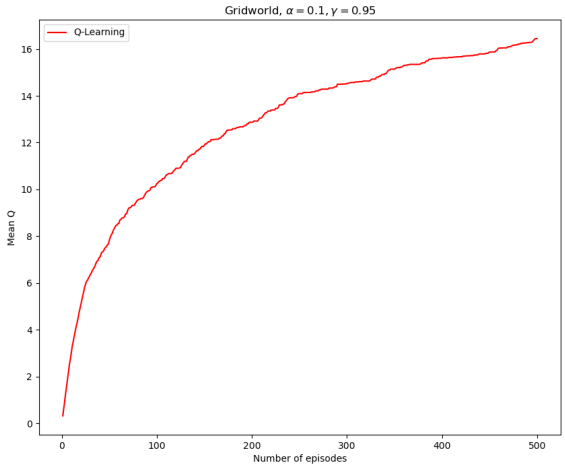

Fig. 11. Learning Curve - Mean Value Function vs. Number of Episodes

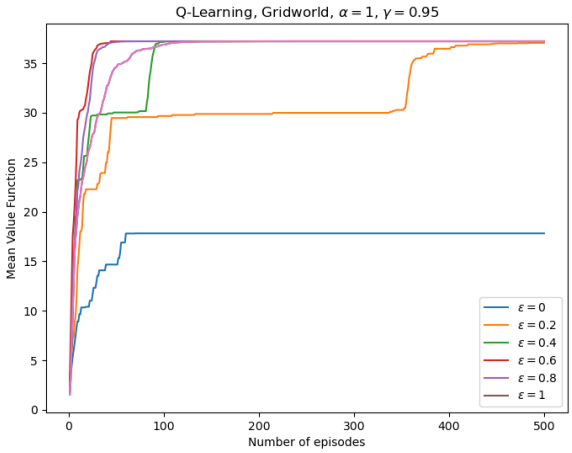

Fig. 13. Learning Curve for Various (ϵ)

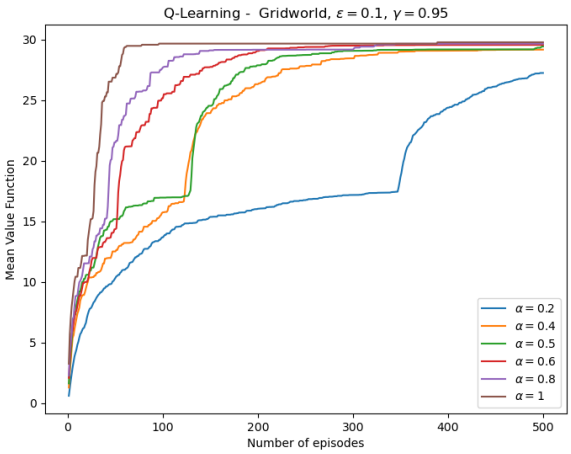

Fig. 12. Learning Curve for Various Learning Rate (α)

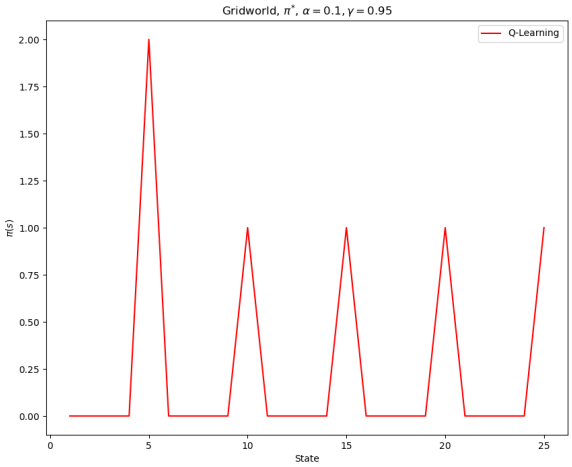

Fig. 14. Action-Value Function vs. Number of Episodes

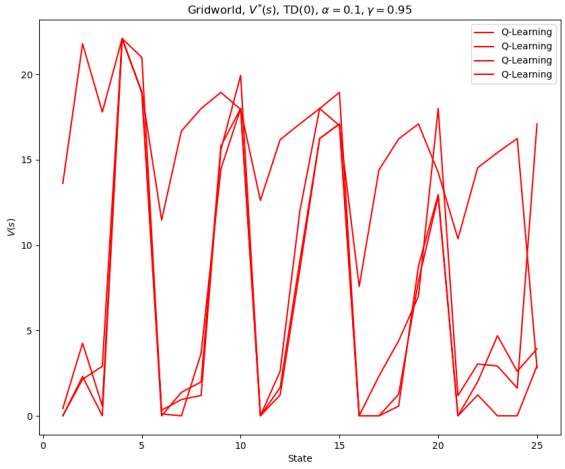

Fig. 15. State-Value Function vs. Number of Episodes

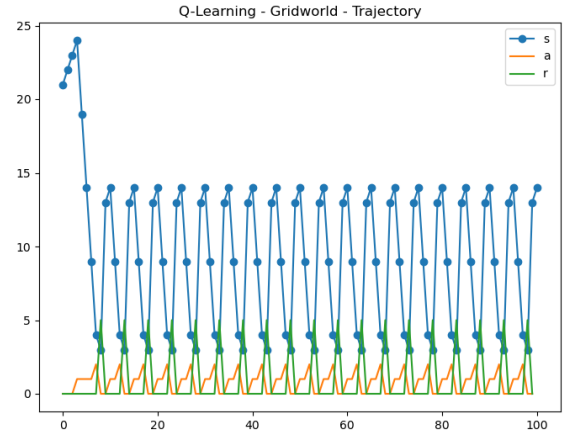

Fig. 16. Policy vs. Trajectory

III. Simple Pendulum

A. SARSA

TABLE III
HYPERPARAMETER VALUES

Hyperparameter	Values
Gamma (γ)	0.95
Epsilon (ϵ)	0 - 1
Alpha (α)	0.2 - 1
Iterations	100
Episodes	500
Threshold (θ)	1e-8

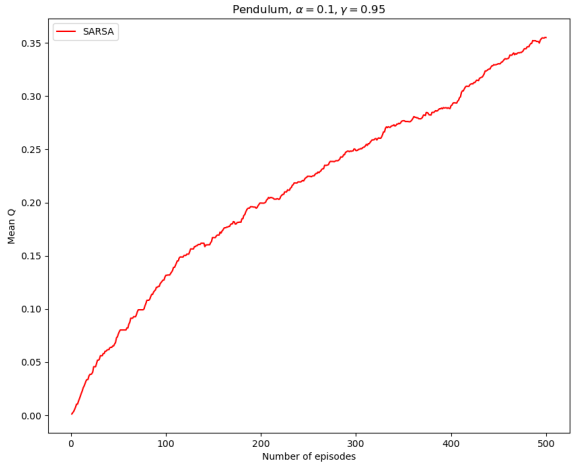

Fig. 17. Learning Curve - Mean Value Function vs. Number of Episodes

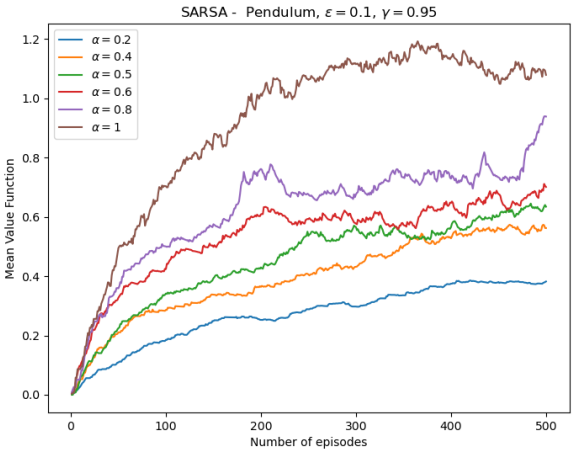

Fig. 18. Learning Curve for Various Learning Rate (α)

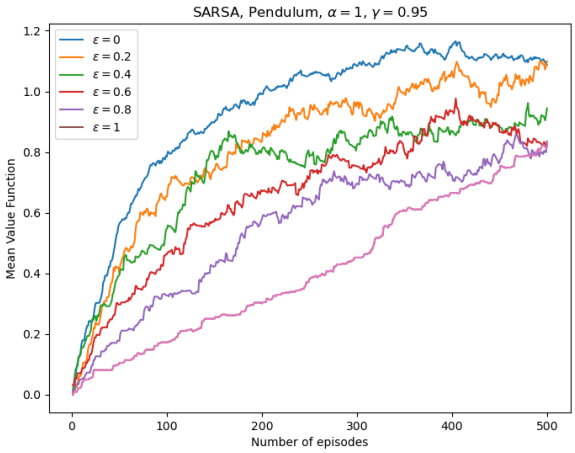

Fig. 19. Learning Curve for Various (ϵ)

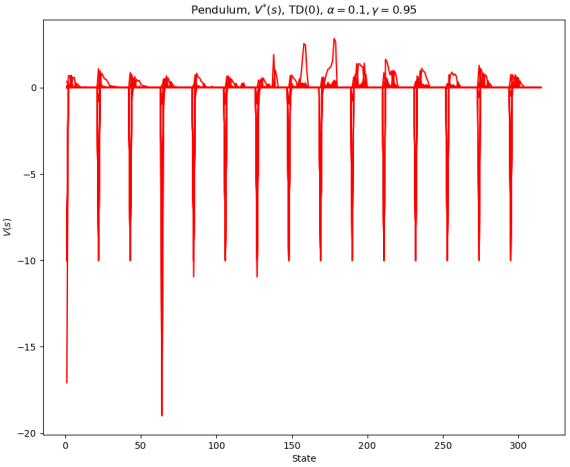

Fig. 21. State-Value Function vs. Number of Episodes

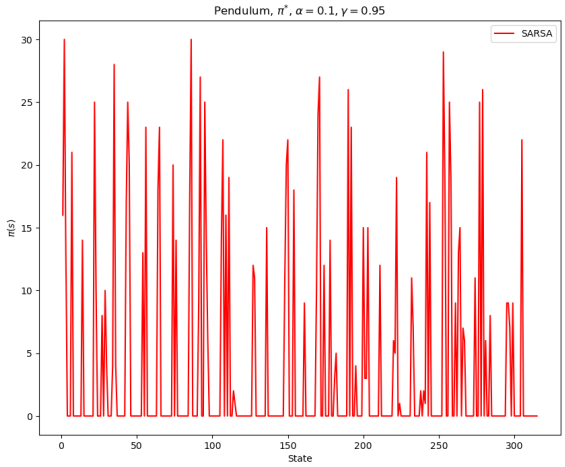

Fig. 20. Action-Value Function vs. Number of Episodes

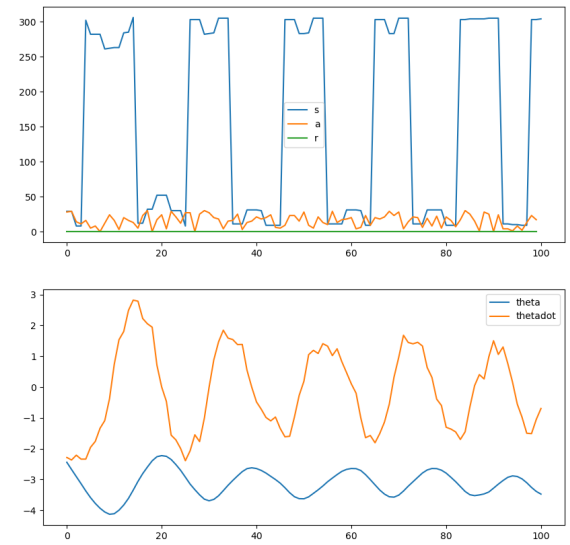

Fig. 22. Policy vs. Trajectory

B. Q-Learning

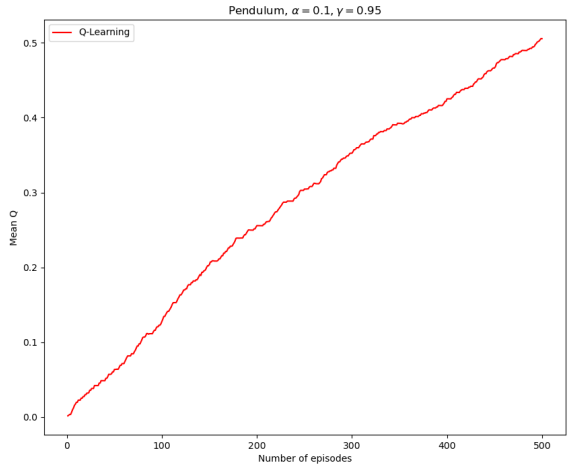

Fig. 23. Learning Curve - Mean Value Function vs. Number of Episodes

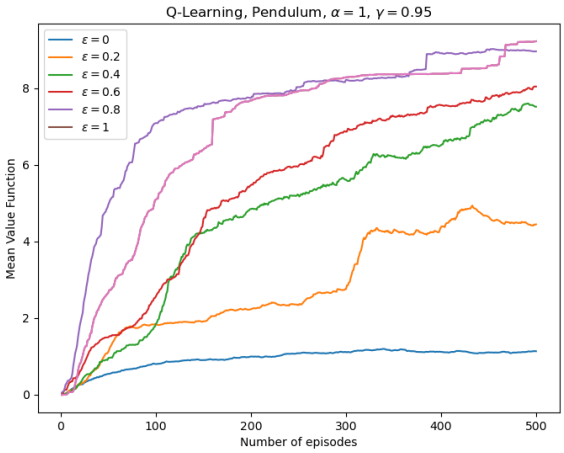

Fig. 25. Learning Curve for Various (ϵ)

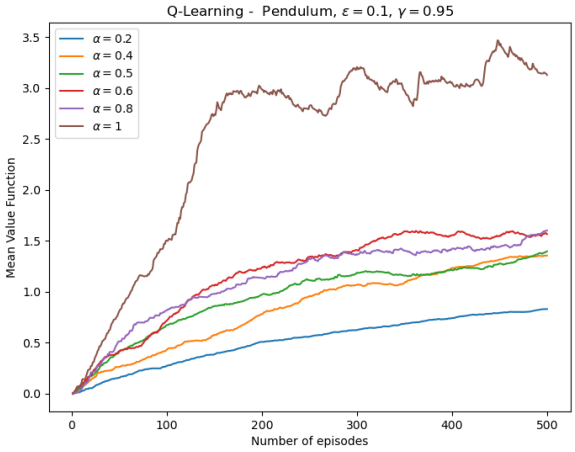

Fig. 24. Learning Curve for Various Learning Rate (α)

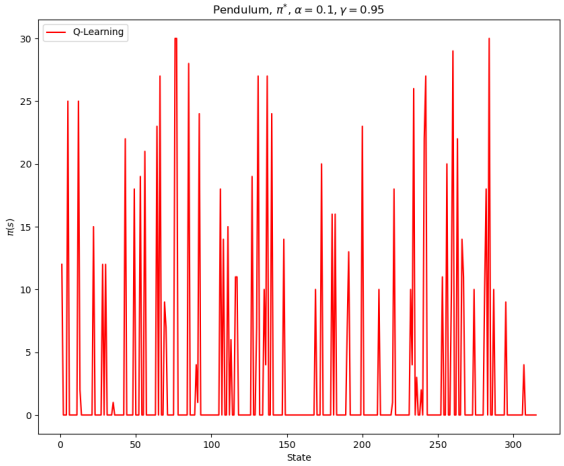

Fig. 26. Action-Value Function vs. Number of Episodes

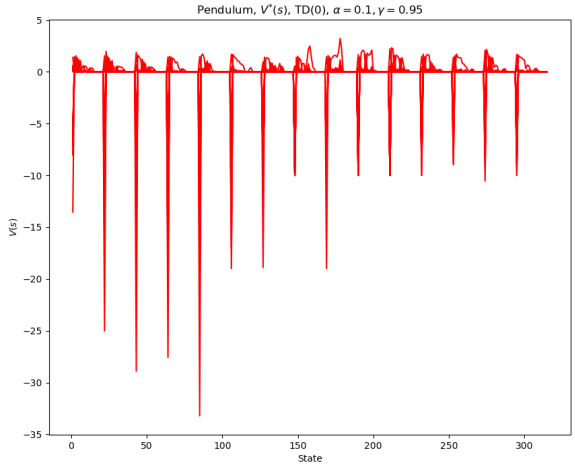

Fig. 27. State-Value Function vs. Number of Episodes

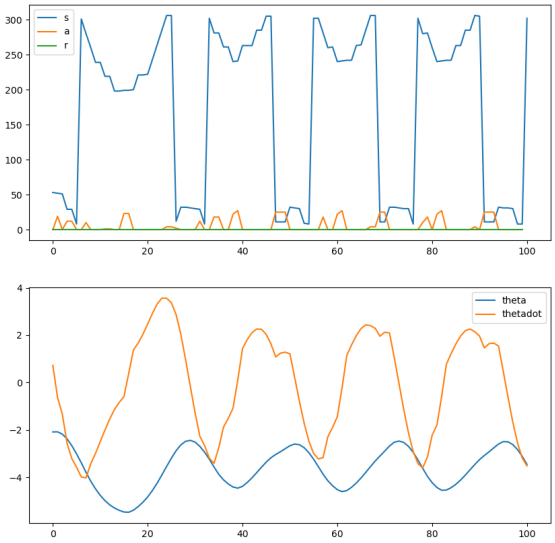

Fig. 28. Policy vs. Trajectory

REFERENCES

[1] Sutton, R. S., and Barto, A. G. Reinforcement learning: An introduction (2nd ed.). The MIT Press. (2018)