Reinformcement Learning Theory Learning

Yue Wang

# Introduction

RL theory is very long history and these days

* Bullet point one
* Bullet point two

1. Numbered list item one
2. Numbered list item two

# Notation and Formatting

some notations:

|  |  |
| --- | --- |
|  | the state space |
|  | the action space |
|  |  |
|  | the state at time t, actual |
|  | the state |
|  |  |
|  | the reward at time t, actual |
|  | the reward from stat s take action to stat ; |
|  |  |
|  | the value of stat s at k iteration |
|  | the value function at time k iteration |
|  |  |
|  | the probability of transfer to when given the current stat s and action a |
|  | the probability of transfer to when given the current stat s and policy |
|  |  |
|  | the policy |
|  | the probability of take action a given the current stat s under policy |

# RL algorithms

# RL convergence theory

## counterexample

There are many that shows the RL algorithms may not convergence even divergence under some conditions

In the authors give an intuitive conclusion about when these algorithms will divergence :

*The danger of instability and divergence arises whenever we combine three things:*

1. training on a distribution of trainsition other than that naturally generated by the process whose expectation is being estimated(e.g. off-policy learning)
2. scalable function approximation (e.g. linear semi-gradient)
3. bootstrapping (eg, DP,TD learning)

There are many counter example, follows are some of that.

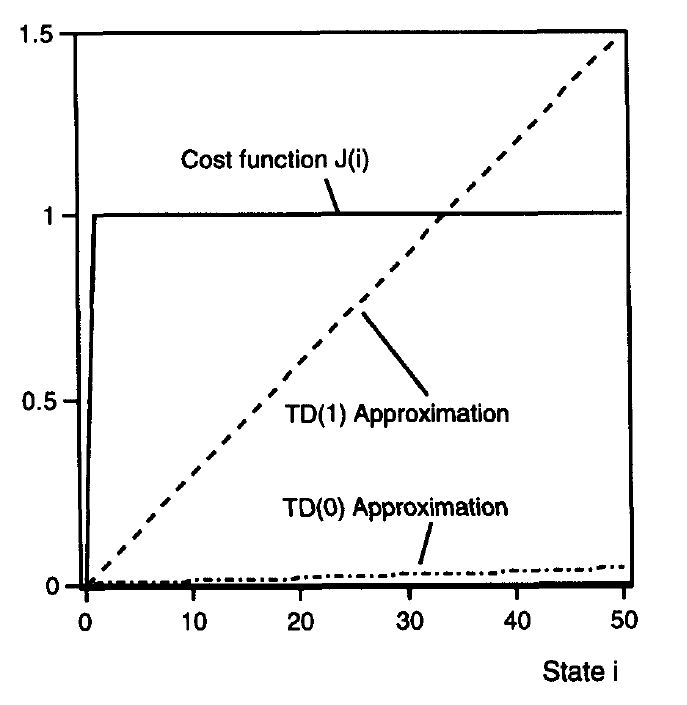
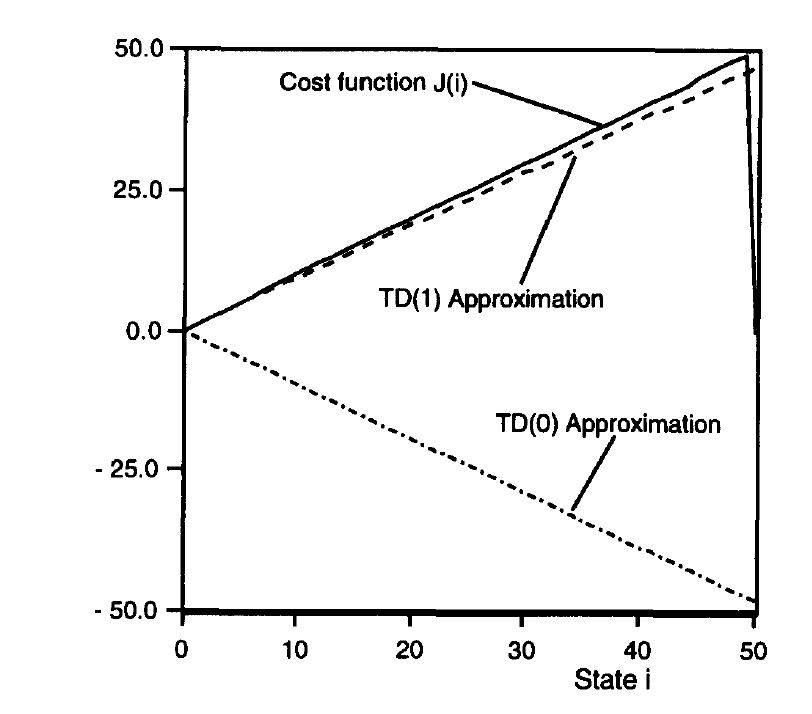
### counterexample1

In the author gives an example to show that the TD(0) algorithms with linear function approximation will diverge.

### conberexample2

In the author gives an example to show that the **TD() algorithm with linear function approximation** convergence to a very poor approximation to the cost function.

As showed in [bertsekas]

### counterexample3

In the author gives an example to show that the value iteration with linear function approximation use off policy may diverge

## convergence for look up table

### policy iteration with look up table

The proof of policy iteration :

First, prove the monotonicity of policy improvement

Second, it is obvious that the number of policy is finite

### value iteration with look up table

The proof of value iteration:

### Q-learning with look up table

### SARSA iteration with look up table

## convergence for linear function approximation

### linear function approximation with prediction problem

### linear function approximation with control problem

## convergence for nonlinear function approximation

### nonlinear function approximation with prediction problem

### nonlinear function approximation with prediction problem