CS489 Assignment 2 Report

517030910214 Hongzhou Liu

0. Introduction

In this assignment, we are required to implement Monte-Carlo (MC) Learning and Temporal-Difference (TD) Learning. These are two methods that can evaluate a given policy directly by learning from episodes of experience without knowledge of MDP model. We will still implement these methods based on GridWorld.

Environment:

- Ubuntu 18.04 LTS
- Python 3.7.4

We've already calculated true value of the given policy in last assignment. We consider it as a baseline.

```
Iterative Policy Evaluation:
-18.17  0.00   -29.22  -44.06  -51.56  -54.68
-32.34  -30.17  -39.60  -47.41  -51.93  -53.80
-44.68  -44.74  -47.58  -50.06  -50.96  -50.79
-52.97  -52.51  -51.95  -50.27  -47.05  -43.61
-57.71  -56.38  -53.44  -48.01  -39.38  -29.00
-59.79  -57.86  -53.42  -44.96  -29.45  0.00
```

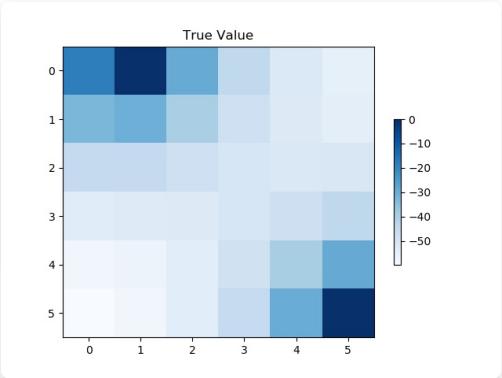

Fig.1 Baseline

We can also visualize

1. Monte-Carlo Learning

1.1 Implementation

1.1.1 First Visit MC

1.1.2 Every Visit MC

1.2 Result

2. Temporal-Difference Learning

2.1 Implementation

2.2 Result

3. Summary