

The Homework of CS285

Deep Reinforcement Learning

Ling Zhang

2025 年 8 月 10 日

目录

1 Homework 1	1
1.1 Analysis	1
1.1.1 Part A	1
1.1.2 Part B	2
1.2 Editing Coding	2
1.2.1 Part A	2
1.2.2 Part B	2
1.3 Discussion	2
2 Homework 2	4
2.1 Introduction	4
2.2 Problem 1	4
2.3 Problem 2	4
3 Homework 3	5
3.1 Introduction	5
3.2 Problem 1	5
3.3 Problem 2	5
4 Homework 4	6
4.1 Introduction	6
4.2 Problem 1	6
4.3 Problem 2	6
5 Homework 5	7
5.1 Introduction	7
5.2 Problem 1	7
5.3 Problem 2	7

Chapter 1: Homework 1

1.1 Analysis

1.1.1 Part A

这个作业相当于是 slide 里条件的弱化版本，slides 里的条件是每个状态不等于专家状态的概率都为 ϵ ，这里只是期望小于 ϵ 。

假设如下条件成立：

$$\mathbb{E}_{p_{\pi^*}(s)} [\pi_\theta(a \neq \pi^*(s) \mid s)] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{p_{\pi^*}(s_t)} [\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)] \leq \epsilon \quad (1.1)$$

在 t 时刻， s_t 的状态分布为：

$$p_\theta(s_t) = (1 - \Pr[\cup_{t'=1}^t \pi_\theta(a_{t'} \neq \pi^*(s_{t'}) \mid s_{t'})]) p_{\pi^*}(s_t) + \Pr[\cup_{t'=1}^t \pi_\theta(a_{t'} \neq \pi^*(s_{t'}) \mid s_{t'})] p_{\text{mistake}}(s_t) \quad (1.2)$$

两边同时减去 $p_{\pi^*}(s_t)$ ，得到：

$$\begin{aligned} |p_\theta(s_t) - p_{\pi^*}(s_t)| &= \Pr[\cup_{t'=1}^t (\pi_\theta(a_{t'} \neq \pi^*(s_{t'}) \mid s_{t'}))] \cdot |p_{\text{mistake}}(s_t) - p_{\pi^*}(s_t)| \\ &\leq 2 \sum_{t=1}^T (\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)) \end{aligned} \quad (1.3)$$

所以：

$$\begin{aligned} \sum_{s_t} |p_\theta(s_t) - p_{\pi^*}(s_t)| &\leq 2 \sum_{t=1}^T \sum_{s_t} p_{\pi^*}(s_t) (\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)) \\ &= 2 \sum_{t=1}^T \mathbb{E}_{p_{\pi^*}(s_t)} [\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)] \\ &= 2T\epsilon \end{aligned} \quad (1.4)$$

得证。

1.1.2 Part B

当奖励函数只与最后一个状态相关时，假设 $J(\pi^*)$ 为专家策略的期望奖励， $J(\pi_\theta)$ 为当前策略的期望奖励。

$$\begin{aligned}
 J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^T (E_{p_{\pi^*}(s_t)} r(s_t) - E_{p_{\pi_\theta}(s_t)} r(s_t)) r(s_t) \\
 &= \sum_{t=1}^T \sum_{s_t} (p_{\pi^*}(s_t) r(s_t) - p_{\pi_\theta}(s_t) r(s_t)) \\
 &= \sum_{s_t} (p_{\pi^*}(s_t) r(s_t) - p_{\pi_\theta}(s_t) r(s_t)) \\
 &\leq 2\epsilon T R_{\max}
 \end{aligned} \tag{1.5}$$

所以：

$$J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\epsilon) \tag{1.6}$$

当为任意奖励时

$$\begin{aligned}
 J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^T (E_{p_{\pi^*}(s_t)} r(s_t) - E_{p_{\pi_\theta}(s_t)} r(s_t)) r(s_t) \\
 &= \sum_{t=1}^T \sum_{s_t} (p_{\pi^*}(s_t) r(s_t) - p_{\pi_\theta}(s_t) r(s_t)) \\
 &\leq 2\epsilon T^2 R_{\max}
 \end{aligned} \tag{1.7}$$

所以：

$$J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\epsilon) \tag{1.8}$$

1.2 Editing Coding

1.2.1 Part A

1.2.2 Part B

1.3 Discussion

Your solution here.

表 1.1: Part 3.1: Behavioral Cloning (BC) 结果表。报告两个任务（一个达到至少 30% 专家性能，一个未达到）。表中为多条 rollout 的平均回报与标准差。公平对比细节（网络结构： $n_layers = ?$, $size = ?$ ；训练： $steps/iter = ?$, $n_iter = ?$ ；专家数据量： $?$ ；评估参数： $ep_len = ?$, $eval_batch_size = ?$ ）请在本 caption 中注明。

Environment	BC Mean Return	BC Std Return	Expert Mean Return	% of Expert
Ant-v4	4786.60	54.70	4681.89	102.2%
HalfCheetah-v4				

Chapter 2: Homework 2

2.1 Introduction

This is the second homework assignment for CS285.

2.2 Problem 1

Your solution here.

2.3 Problem 2

Your solution here.

Chapter 3: Homework 3

3.1 Introduction

This is the third homework assignment for CS285.

3.2 Problem 1

Your solution here.

3.3 Problem 2

Your solution here.

Chapter 4: Homework 4

4.1 Introduction

This is the fourth homework assignment for CS285.

4.2 Problem 1

Your solution here.

4.3 Problem 2

Your solution here.

Chapter 5: Homework 5

5.1 Introduction

This is the fifth homework assignment for CS285.

5.2 Problem 1

Your solution here.

5.3 Problem 2

Your solution here.