# CS 285 hw1

oliver song

October 2023

## 1

### 1.1

Suppose the probability $\epsilon_t$ of making a mistake at time t is given by

$$\pi_\theta(a \neq \pi^*(s_t)|s_t) = \epsilon_t$$

From the given condition we can show that

$$\frac{1}{T}\sum_{i=1}^{T}\epsilon_i \leq \epsilon$$

We can denote the state distribution at time $t$ as $p_{\pi_\theta}(s_t)$, which can be written as the sum of probability of getting the correct state $p_{\pi^*}(s_t)$ and of getting the wrong state $p_{mistake}(s_t)$:

$$p_{\pi_\theta}(s_t) = \prod_{i=1}^{t}(1-\epsilon_i)\cdot p_{\pi^*}(s_t) + (1-\prod_{i=1}^{t}(1-\epsilon_i))\cdot p_{mistake}(s_t)$$

Applying some mathematical transformation to the equation yields

$$\begin{aligned} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| &= (1-\prod_{i=1}^{t}(1-\epsilon_i))|p_{\pi^*}(s_t) - p_{mistake}(s_t)| \\ &\leq 2\sum_{1}^{t}\epsilon_t \\ &\leq 2\epsilon T \end{aligned} \tag{1}$$

### 1.2

#### 1.2.1

From 1.1 we know that the bias of the last state has dimension $O(T\epsilon)$. Since the reward is only related to the final state, the bias of return should also have dimension $O(T\epsilon)$

### 1.2.2

From the conclusion in 1.1 we can also know that at every state $s_t$, the distribution bias is bounded by $2T\epsilon$. According to the definition of $J(\pi)$,

$$J(\pi^*) - J(\pi_\theta) \leq 2T^2 \epsilon R_{max}$$

which has the dimension $O(T^2\epsilon)$