Section 2 (Part 2): Behavior Cloning Regret Proof

1 Setup: Behavior Cloning and Distributional Shift

The Problem

In behavior cloning, we train a policy $\pi_\theta(a_t|s_t)$ by imitating an expert policy $\pi^*(s_t)$ using supervised learning on demonstration data.

Training objective:

$$\max_\theta \mathbb{E}_{s_t \sim p_{\text{data}}(s_t)} [\log \pi_\theta(a_t = \pi^*(s_t)|s_t)]$$

The Issue: We train on states from $p_{\text{data}}(s_t)$ (expert's state distribution), but at test time the learned policy induces its own distribution $p_{\pi_\theta}(s_t)$.

Distributional shift:

$$p_{\text{data}}(s_t) \neq p_{\pi_\theta}(s_t)$$

Question: How bad can this distributional shift be? Can we quantify the error?

2 Defining the Cost Function

To analyze the problem, we define a simple cost function that measures "mistakes":

Cost function:

$$c(s_t, a_t) = \begin{cases} 0 & \text{if } a_t = \pi^*(s_t) \\ 1 & \text{otherwise} \end{cases}$$

The expected cost at a single time step, given state s_t:

$$\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}[c(s_t, a_t)] = \sum_{a_t} \pi_\theta(a_t|s_t) \cdot c(s_t, a_t) = \pi_\theta(a_t \neq \pi^*(s_t)|s_t)$$

This is simply the probability of making a mistake at state s_t.

3 The Key Assumption

We assume that our learned policy makes mistakes with probability at most ϵ *on states from the training distribution*:

Assumption:

$$\mathbb{E}_{s_t \sim p_{\text{train}}(s_t)} [\pi_\theta(a_t \neq \pi^*(s_t)|s_t)] \leq \epsilon$$

Interpretation: If we sample a state from where the expert went, we are unlikely (probability $\leq \epsilon$) to make a mistake.

Goal: Bound the *total expected number of mistakes* over a trajectory of length H:

$$\sum_{t=1}^{H} \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t),\, s_t \sim p_{\pi_\theta}(s_t)} \left[c(s_t, a_t)\right]$$

Note: The states s_t are sampled from $p_{\pi_\theta}(s_t)$ — the distribution induced by *running our policy*, not from training data!

4 Step 1: State Distribution Decomposition

Key insight: We can decompose the state distribution under π_θ into two cases:

1. We made **no mistakes** so far → we're still on the training distribution
2. We made **at least one mistake** → we're on some other distribution

State distribution decomposition:

$$p_{\pi_\theta}(s_t) = (1-\epsilon)^t\, p_{\text{train}}(s_t) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}(s_t)$$

Where:

- $(1-\epsilon)^t$ = probability of making no mistakes for t steps
- $p_{\text{train}}(s_t)$ = state distribution if we always acted like the expert
- $p_{\text{mistake}}(s_t)$ = some (unknown) distribution we land in after a mistake

5 Step 2: Bounding the TV Divergence

Total Variation Divergence measures how different two distributions are:

$$D_{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)| \leq 1$$

Now we compute the TV divergence between p_{train} and p_{π_θ}:

$$
\begin{aligned}
D_{TV}(p_{\text{train}}, p_{\pi_\theta}) &= \frac{1}{2}\sum_{s_t} |p_{\text{train}}(s_t) - p_{\pi_\theta}(s_t)| \\
&= \frac{1}{2}\sum_{s_t} \left|p_{\text{train}}(s_t) - (1-\epsilon)^t p_{\text{train}}(s_t) - (1-(1-\epsilon)^t) p_{\text{mistake}}(s_t)\right| \\
&= \frac{1}{2}\sum_{s_t} \left|(1-(1-\epsilon)^t) p_{\text{train}}(s_t) - (1-(1-\epsilon)^t) p_{\text{mistake}}(s_t)\right| \\
&= (1-(1-\epsilon)^t) \cdot \frac{1}{2}\sum_{s_t} |p_{\text{train}}(s_t) - p_{\text{mistake}}(s_t)| \\
&\leq (1-(1-\epsilon)^t)
\end{aligned}
$$

Useful inequality: For $\epsilon \in [0, 1]$:

$$(1-\epsilon)^t \geq 1 - \epsilon t \quad \Longrightarrow \quad 1-(1-\epsilon)^t \leq \epsilon t$$

Proof sketch: Bernoulli's inequality states $(1+x)^n \geq 1 + nx$ for $x \geq -1$.

Therefore:

$$D_{TV}(p_{\text{train}}, p_{\pi_\theta}) \leq \epsilon t$$

The distributions diverge **linearly** with time.

6 Step 3: Bounding the Total Cost

We want to bound:

$$\sum_{t=1}^{H} \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t),\, s_t \sim p_{\pi_\theta}(s_t)} \left[c(s_t, a_t)\right]$$

Strategy: Add and subtract $p_{\text{train}}(s_t)$ to relate back to our assumption.

$$
\begin{aligned}
&\sum_{t=1}^{H} \mathbb{E}_{s_t \sim p_{\pi_\theta}(s_t)} \left[\mathbb{E}_{a_t \sim \pi_\theta}[c(s_t, a_t)]\right] \\
&= \sum_{t=1}^{H} \sum_{s_t} p_{\pi_\theta}(s_t)\, \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}[c(s_t, a_t)]
\end{aligned}
$$

Add and subtract $p_{\text{train}}(s_t)$:

$$= \sum_{t=1}^{H} \sum_{s_t} \left(p_{\text{train}}(s_t) + p_{\pi_\theta}(s_t) - p_{\text{train}}(s_t)\right) \mathbb{E}_{a_t}[c(s_t, a_t)]$$

$$= \sum_{t=1}^{H} \left[\underbrace{\sum_{s_t} p_{\text{train}}(s_t) \mathbb{E}_{a_t}[c(s_t, a_t)]}_{\text{Term A}} + \underbrace{\sum_{s_t} (p_{\pi_\theta}(s_t) - p_{\text{train}}(s_t)) \mathbb{E}_{a_t}[c(s_t, a_t)]}_{\text{Term B}} \right]$$

7 Step 4: Bounding Each Term

Term A: Expected cost under training distribution.

By our assumption:

$$\text{Term A} = \mathbb{E}_{s_t \sim p_{\text{train}}(s_t)}[\pi_\theta(a_t \neq \pi^*(s_t)|s_t)] \leq \epsilon$$

Term B: Extra cost due to distributional shift.

$$\text{Term B} = \sum_{s_t} (p_{\pi_\theta}(s_t) - p_{\text{train}}(s_t)) \mathbb{E}_{a_t}[c(s_t, a_t)]$$

$$\leq \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\text{train}}(s_t)| \cdot \underbrace{\mathbb{E}_{a_t}[c(s_t, a_t)]}_{\leq 1}$$

Since the cost is bounded by 1:

$$\text{Term B} \leq \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\text{train}}(s_t)| = 2\, D_{TV}(p_{\pi_\theta}, p_{\text{train}}) \leq 2\epsilon t$$

8 Step 5: Final Result

Combining the bounds:

$$\sum_{t=1}^{H} \mathbb{E}[c(s_t, a_t)] \leq \sum_{t=1}^{H} (\epsilon + 2\epsilon t)$$

$$= \epsilon H + 2\epsilon \sum_{t=1}^{H} t$$

$$= \epsilon H + 2\epsilon \cdot \frac{H(H+1)}{2}$$

$$= \epsilon H + \epsilon H(H+1)$$

$$= \epsilon H(1 + H + 1)$$

$$= O(\epsilon H^2)$$

Main Result: The expected total number of mistakes scales as

$$\boxed{O(\epsilon H^2)}$$

Quadratic in the horizon H.

9 Interpretation and Takeaways

Observation	Implication
Error is $O(\epsilon H^2)$	Small mistakes **compound** over time
Quadratic in H	Long horizons are **much harder** than short ones
Linear in ϵ	Reducing per-step error helps, but doesn't fix the H^2

Contrast with i.i.d. supervised learning:

In standard supervised learning, if per-sample error is ϵ, total error over H samples is $O(\epsilon H)$ (linear). In behavior cloning, it's $O(\epsilon H^2)$ (quadratic) because samples are **not** i.i.d. — current actions affect future states.

What can we do?

- **DAgger:** Collect data from p_{π_θ} instead of p_{data} to match distributions