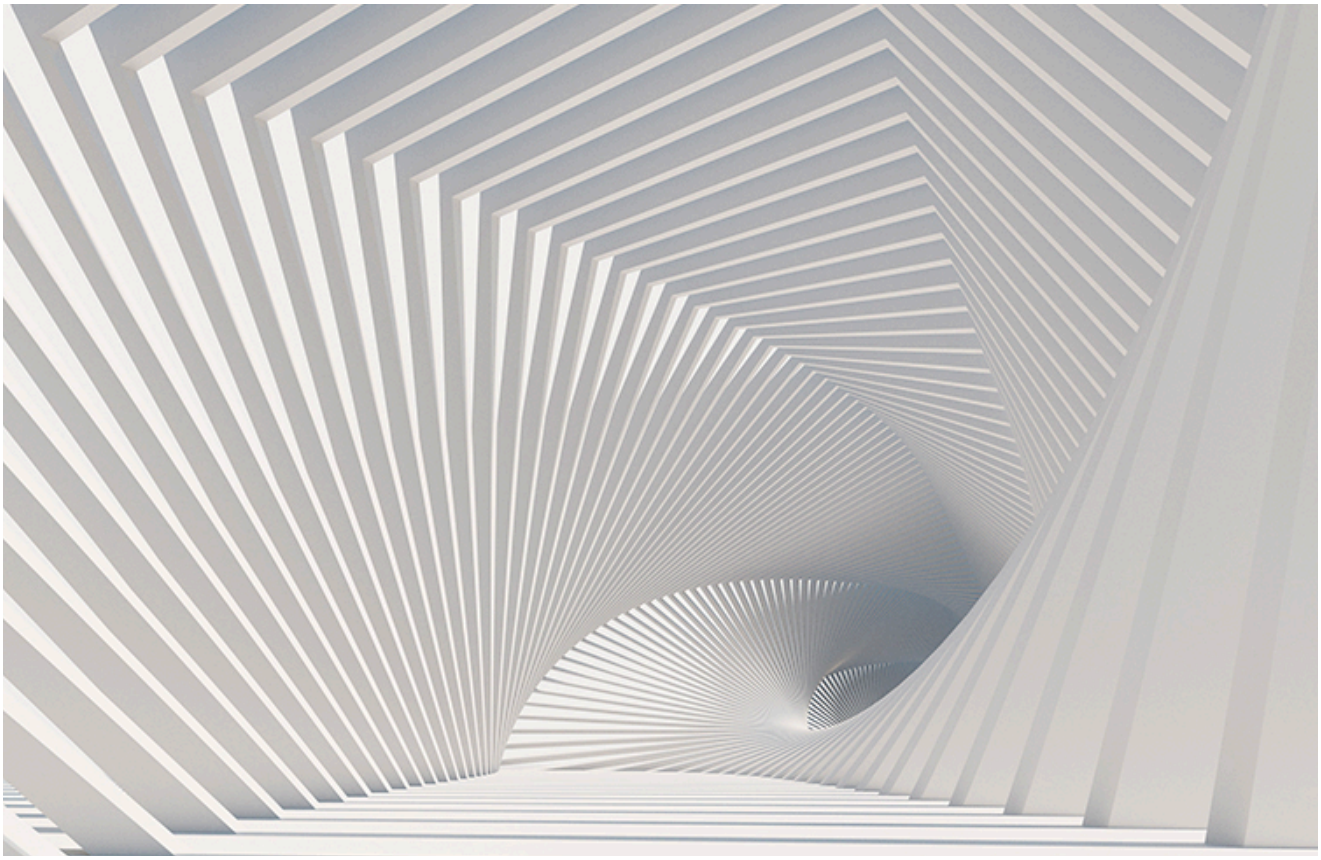
CS294 - DEEP RL



HW1. Behavior Cloning

There are quite a lot of materials to cover, we need to split the work to daily routine.

Found: In terms of the clarity of explanation, the slide merely guides your intuition, but the rigor of the displayed equation is horrible — however, it does the job of referencing people to the correct source. For me, the paper '***A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning***' does a much better job at explaining even the basic concept in a flawless way.

The key question, (even the lecture does not explain this well, the slides also not doing good job) is why does

$$\pi_\theta(a \neq \pi^*(s) \,|\, s) \leq \epsilon \,.\, \forall s \in \mathscr{D}_{train}$$

Implies

$$\mathbb{E}[\sum_t c(s_t, a_t)] \leq \epsilon T + (1-\epsilon)(\epsilon(T-1) + (1-\epsilon(\dots))) = \mathcal{O}(\epsilon T)$$

Solved:

The proof actually resides in a index-3 assemble reference. The supplementary proof of the paper '***Efficient Reductions for Imitation Learning***' (Stephen Ross)

π_θ, π^* — learned policy, demo (expert) policy (assumed to be deterministic !)

$C(s,a)$ — Immediate cost of doing action a in state s

$C_\pi(s) := \int C(s,a)\pi(a \mid s)da = \mathbb{E}_{a\sim\pi}[C(s,a)]$—expected immediate cost of performing policy in s

d_π^i — state distribution at time step i if we follow policy π from the initial time step

$d_\pi = \frac{1}{T}\sum_{i=1}^{T} d_\pi^i$ — average state visiting distribution (more often we use a discounted version of this)

$e(s,a) = I(a \neq \pi^*(s))$ — 0-1 loss of executing action a in state s

$e_\pi = \mathbb{E}_{a\sim\pi}(e(s,a))$ just like the definition of expected immediate cost, except for the 0-1 loss case

$J(\pi) = T\mathbb{E}_{s\sim d_\pi}[C_\pi(s)] = \mathbb{E}[\sum_{t=1}^{T} C_\pi(s_t)]$ — expected T-step cost of executing policy π

$\mathcal{R}_\prod(\pi) = J(\pi) - \min_{\pi'\in\prod} J(\pi')$ — regret of policy w.r.t the best policy in a particular policy class $\prod$

Theorem 2.1.

Let $\hat{\pi}$ be such that $\mathbb{E}_{s\sim d_{\pi^*}}[e_{\hat{\pi}}(s)] \leq \epsilon$. Then $J(\hat{\pi}) \leq J(\pi^*) + T^2\epsilon$

Key Missing part (took ~ 3 hours to figure out) from the proof is to show that

$$\mathbb{E}_{s\sim d_t}(C_{\hat{\pi}}(s)) \leq \mathbb{E}_{s\sim d_t}(C_{\pi^*}(s)) + e_t \qquad \text{(eq. 1)}$$

Here we note that by definition

$$e_t := \mathbb{E}_{s\sim d_t}[1 - \hat{\pi}(a = \pi^*(s) \mid s)]$$

In order to show (eq. 1), It suffices to show

$$C_{\hat{\pi}}(s) \leq C_{\pi^*}(s) + \sum_{a\neq\pi^*(s)} \pi(a \mid s)$$

$$C_{\hat{\pi}}(s) = \sum_a C(s,a)\hat{\pi}(a \mid s)$$

$$\leq \sum_{a\neq\pi^*(s)} \hat{\pi}(a \mid s) + \sum_{a=\pi^*(s)} C(a,s)\hat{\pi}(a \mid s)$$

$$\leq \sum_{a\neq\pi^*(s)} \hat{\pi}(a \mid s) + \sum_{a=\pi^*(s)} C(a,s)\mathbb{I}(a = \pi^*(s))$$

$$\leq \sum_{a \neq \pi^*(s)} \hat{\pi}(a \mid s) + \sum_{a=\pi^*(s)} C(a,s)\pi^*(a \mid s)$$
$$\leq \sum_{a \neq \pi^*(s)} \hat{\pi}(a \mid s) + C_{\pi^*}(s)$$

End of proof ...

An Alternative argument is if we just use the 1-0 loss, then this inequality looks redundant. Another useful law is

$$\mathbb{P}(\bigcup_i E_i) \leq \sum_i \mathbb{P}(E_i)$$

Note that the lecture note uses 1-0 loss as the cost function, this would implies a much tighter upper-bound exists, and some of the process above is redundant, albeit the inequality still holds.

— Explain of In-Class reasoning on the 1-0 loss case: (Intuition adopted to speeds up proving)
Here is a few *mentality trick* to get there quickly
—

Corollary 1. Assume $\pi_\theta(a \neq \pi^*(s) \mid s) \leq \epsilon, \forall s \in \mathcal{S}$, we have

$$\mathbb{E}[\sum_t c(s_t, a_t)] = \mathcal{O}(\epsilon T^2)$$

Proof:

We split the expectation into summation of conditional expectation: when the first $t-1$ actions from learned policy are different from expert actions, and the t^{th} action matches with the expert action.

$$\mathbb{E}[\sum_t^T c(s_t, a_t)] = \mathbb{E}[\sum_t^T c(s_t, a_t) \mid c(s_1, a_1) = 1]\mathbb{P}(c(s_1, a_1) = 1)$$
$$+\mathbb{E}[\sum_{t=1}^T c(s_t, a_t) \mid c(s_1, a_1) = 0, c(s_2, a_2) = 1]\mathbb{P}(c(s_1, a_1) = 0, c(s_2, a_2) = 1) + \dots$$

Here is a trick, we need only an upper-bound, and we do not want any exact calculation here, so basically

$$\mathbb{P}(\dots, c(s_t, a_t) = 1) \leq \mathbb{P}(c(s_t, a_t) = 1) \leq \epsilon$$

(Marginal distribution always bigger than the joint distribution)

$$\sum_{t=1}^T c(s_t, a_t) \leq T$$

(By definition of cost function, it ranges between [0, 1])

Above two argument tells me the corollary holds. Completed.

Now we look at the HW problem. BTW, the argument provided in-class is wrong and is quite misleading, — although the intuition is probably fine there... Next time we get stuck on a CS course, try an alternative approach as the lecture is not rigid.

Homework Q1.1

Assume random demo policy $\pi^*(a \mid s)$, prove

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*(s_t)}}[\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)] \leq \epsilon \implies \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2T\epsilon$$

Intuition: Closeness in policy leads to closeness in trajectories — average visitation policy closeness leads to linearly-bad bounds on trajectory difference.

(eq.1)

$$\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) = \sum_{a_t \neq \pi^*(s_t)} \pi_\theta(a_t) = \sum_{a_t} \pi_\theta(a_t \mid s_t)(1 - \pi^*(s_t)) \geq \sum_{a_t} (\pi_\theta(a_t \mid s_t) - \pi^*(a_t \mid s_t))$$

This equations shows how the likelihood of difference between policy relates to prob density diff

$$\mathbb{E}_{p_{\pi^*(s_t)}}[\sum_{a_t} (\pi_\theta(a_t \mid s_t) - \pi^*(a_t \mid s_t))] \leq \mathbb{E}_{p_{\pi^*(s_t)}}[\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)] \leq T\epsilon$$

Now we use induction to conclude, since the initial distribution $p(s_1)$ is decided by the environment (and not the policy) the $t = 1$ case naturally holds.

Assume the conclusion holds for t, we will prove the case for $t + 1$

$$\sum_{s_{t+1}} |p_{\pi_\theta}(s_{t+1}) - p_{\pi^*}(s_{t+1})| = \sum_{s_{t+1}} |\sum_{s_t, a_t} \left(p_{\pi_\theta}(s_t)\pi_\theta(a_t \mid s_t)T(s_{t+1} \mid s_t, a_t) - p_{\pi^*}(s_t)\pi^*(a_t \mid s_t)T(s_{t+1} \mid s_t, a_t)\right)|$$

Here we can apply a classic decomposition

$$a_1 b_1 - a_2 b_2 = (a_1 - a_2)b_1 + (b_1 - b_2)a_2$$

To obtain:

$$p_{\pi_\theta}(s_t)\pi_\theta(a_t \mid s_t) - p_{\pi^*}(s_t)\pi^*(a_t \mid s_t) = (p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))\pi_\theta(a_t \mid s_t) + p_{\pi^*}(s_t)(\pi_\theta(a_t \mid s_t) - \pi^*(a_t \mid s_t))$$

Now we decompose the full equation

$$\sum_{s_{t+1}} |p_{\pi_\theta}(s_{t+1}) - p_{\pi^*}(s_{t+1})| = \sum_{s_{t+1}} |\sum_{s_t, a_t} \left(p_{\pi_\theta}(s_t)\pi_\theta(a_t \mid s_t) - p_{\pi^*}(s_t)\pi^*(a_t \mid s_t)\right)T(s_{t+1} \mid s_t, a_t)|$$

$$= \sum_{s_{t+1}} |\sum_{s_t, a_t} \left((p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))\pi_\theta(a_t \mid s_t)\right)T(s_{t+1} \mid s_t, a_t)| \text{ — (term 1)}$$

$$+ \sum_{s_{t+1}} |\sum_{s_t, a_t} \left(p_{\pi^*}(s_t)(\pi_\theta(a_t \mid s_t) - \pi^*(a_t \mid s_t))\right)T(s_{t+1} \mid s_t, a_t)| \text{ — (term 2)}$$

For term 1, we have

$$\sum_{s_{t+1}} |\sum_{s_t,a_t} \big((p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))\pi_\theta(a_t \,|\, s_t)\big) T(s_{t+1} \,|\, s_t, a_t)| \le \sum_{s_{t+1}} |\sum_{s_t,a_t} |(p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))| \,\pi_\theta(a_t \,|\, s_t) T(s_{t+1} \,|\, s_t, a_t)|$$

$$\le \sum_{s_t,a_t} |(p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))| \,\pi_\theta(a_t \,|\, s_t) \le \sum_{s_t} |(p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t))| \le T\epsilon$$

For term 2, we have

$$\sum_{s_{t+1}} \sum_{s_t,a_t} \big(p_{\pi^*}(s_t)(\pi_\theta(a_t \,|\, s_t) - \pi^*(a_t s_t))\big) T(s_{t+1} \,|\, s_t, a_t) \le \sum_{s_t,a_t} p_{\pi^*}(s_t) \,|\, \pi_\theta(a_t \,|\, s_t) - \pi^*(a_t s_t) \,| \le T\epsilon$$

As a result, we have

$$\sum_{s_{t+1}} |\, p_{\pi_\theta}(s_{t+1}) - p_{\pi^*}(s_{t+1}) \,| \le 2\epsilon T$$

End of proof.

Q 1.2 & Q 1.3 is trivial

Assume upper-bound (normalization) on the reward values, $r(s_t) \le R_{max} \,\forall s_t$ as well as the closeness assumption on the policy

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*(s_t)}}[\pi_\theta(a_t \neq \pi^*(s_t) \,|\, s_t)] \le \epsilon$$

Then the reward value difference between the trained policy & target policy has boundary

$$J(\pi_\theta) - J(\pi^*) = \sum_t \sum_{s_t} r(s_t) \Big(p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t) \Big) \le T R_{max} T\epsilon$$

A case that Q1.2 tries to make is that when the reward is only provided by the final state, or sparse reward is available only, in that case the difference between the value function of the two policy is smaller (as in the difference is bounded by some terms which grows linearly in time, instead of quadratic in time)

End of proof.

I guess the important lesson here, is just that the condition, which underlines the training approach for the induced algorithm is to 'match' trained policy with the demo policy on those demo trajectories, since then the error / mistakes incurred for actual experiment (trajectories run with trained policy) will be quadratic in time dimension (in terms of the value function, densely distributed in state space).

I wonder if this fundamental flaw can be addressed at all, by any sort of behavior cloning algorithm, a practical implementation demonstrated in class is the Nvidia's auto pilot attempt, which train also with augmented 'weird' driving scene, designed to steer back from errored direction.