Imitation Learning Notes

1

1.1

Algorithm 1 DAgger (dataset aggregation)

while not satisfactory performance achieved **do**
  Train the policy π_{data} based on the D
  Collect new trajectories D' using π_{data}
  Ask human to label the new trajectory D' with action a_t
  Aggregate the new trajectories to the dataset $D = D \cup D'$
end while

Proper choice of reward function for imitation learning:

$$r(s, a) = \log p(a = \pi^*(s)|s) \tag{1}$$

or:

$$c(s, a) = \begin{cases} 0 & \text{if } a = \pi^*(s) \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

assume $\pi_\theta(a \neq \pi^*(s)|s) \leq \epsilon$, for all $s \in \mathcal{D}_{\text{train}}$

$$E[\sum_t c(s_t, a_t)] \leq \epsilon T + (1 - \epsilon)(\epsilon(T - 1) + (1 - \epsilon)(...)) \tag{3}$$

which has T terms and each of them is in $O(\epsilon T)$, resulting in a total complexity of $O(\epsilon T^2)$, which is an indication of why naive behavior cloning might be a bad behavior.

1.2

More general analysis: for $s \in p_{\text{train}}(s)$, actually enough for $E_{p_{\text{train}}(s)}[\pi_\theta(a \neq \pi^*(s)|s)] \leq \epsilon$

with DAgger, $p_{\text{train}}(s) \to p_\theta(s)$

if $p_{\text{train}}(s) \neq p_\theta(s)$:

$$p_\theta(s_t) = (1 - \epsilon)^t p_{\text{train}}(s_t) + (1 - (1 - \epsilon)^t))p_{\text{mistake}}(s_t) \tag{4}$$

where the first term denotes the case in which we made no mistakes, and the second denotes the case in which we have made some mistakes (some other distributions). Rearranging we get:

$$|p_\theta(s_t) - p_{\text{train}}(s_t)| = (1 - (1 - \epsilon)^t)|p_{\text{mistake}}(s_t) - p_{\text{train}}(s_t)| \leq 2(1 - (1 - \epsilon)^t) \tag{5}$$

useful identity: $(1 - \epsilon)^t \geq 1 - \epsilon t$ for $\epsilon \in [0, 1]$ As a result:

$$|p_\theta(s_t) - p_{\text{train}}(s_t)| = (1 - (1 - \epsilon)^t)|p_{\text{mistake}}(s_t) - p_{\text{train}}(s_t)| \leq 2\epsilon t \tag{6}$$

Further, we have:

$$\sum_t E_{p_\theta}[c_t] = \sum_t \sum_{s_t} p_\theta(s_t)c_t(s_t) \leq \sum_t \sum_{s_t} p_{\text{train}}(s_t)c_t(s_t) + |p_\theta(s_t) - p_{\text{train}}(s_t)|c_{\max} \leq \sum_t \epsilon + 2\epsilon t \leq \epsilon T + 2\epsilon T^2 \tag{7}$$

1.3 Goal-conditioned behavior cloning

for each demo $s_1^i, a_1^i, ..., s_{T-1}^i, a_{T-1}^i, s_T^i$,
maximize $\log \pi_\theta(a_t^i | s_t^i, g = s_T^i)$
Going beyond just imitation:

- start with random policy
- collect data with random policies
- treat this data as "demonstrations" for the goals that were reached
- use this to improve the policy
- repeat