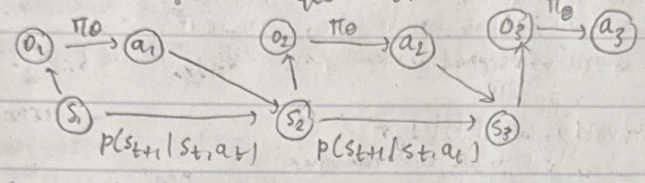
CS 182 Lecture 14: Imitation Learning

Imitation learning = learning to control!

↳ Challenges: variables are NOT i.i.d. (future is affected)] - the issue w/ BC.
goals are abstract, no ground truth labels!
→ i.e. we aim to accomplish a task.
→ we are constantly "controlled" via feedback] - RL to the rescue.

o_t - observation
a_t - action
$\pi_\theta(a_t|o_t)$ - model / policy (timestep)
s_t - state
$\pi_\theta(a_t|s_t)$ - policy (fully observed)

Modeled through a Bayes Net:



s_t obeys Markov property
o_t does not.

Simplest form of IL: existing supervised learning techniques (ex. ConvNet).
↳ "behavioral cloning"
↳ oddly doesn't work in theory, but can work in practicum.

Problem: Compounding error. $p_{\pi_\theta}(o_t)$ deviates gradually from $p_{data}(o_t)$.
aka "distributional shift", "shifts" from true strings to synthetic strings.
→ we also don't know $p(s_{t+1}|s_t, a_t)$. $p_{data}(o_t) \approx p_\theta(o_t)$.
→ mitigate problem thru (1) lot of data, (2) more accurate policy.

Reasons it's hard to mitigate: (1) non-Markovian behavior (fail to fit the expert)
(2) Multimodal behavior

$o_1, \dots, o_t$ → use entire history to share in dependencies. Aka "many-to-one" RNN.

"soft" approximation (averaging effects)
1. Output Gaussian mixture model (learn more about this)
2. Latent variable models
↳ tweak the input, randomly sample vector to represent "latent variables" (these are "guessed", NOT noise!)
→ more: conditional variational autoencoder, (CVAE)
normalizing flow / realNVP.
3. Autoregressive discretization
↳ discretize: one dimension at a time, condition subsequent dim on prev. (seq2seq behavior)

→ use hacks to make BC work (or models with memory).
↳ ex. NVIDIA self-driving approach

Bigger question: Can we "force" $p_{data}(o_t) = p_{\pi_\theta}(o_t)$?
↳ fix $p_{data}(o_t)$ → fix the data, not the policy/problem.

Solution: Dataset Aggregation (DAgger), fixes "distributional shift"
↳ collect training data from $p_{\pi_\theta}(o_t)$, not $p_{data}(o_t)$
↳ to collect labels a_t:
1. train $\pi_\theta(a_t|o_t)$ from human data $\mathcal{D} = \{o_1, a_1, \dots, o_N, a_N\}$ (initial policy)
2. run $\pi_\theta(a_t|o_t)$ to get dataset $\mathcal{D}_\pi = \{o_1, \dots, o_M\}$. ← won't be good, LOL.
3. Ask human to label $\mathcal{D}_\pi$ w/ a_t ("ground truth")
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

main flaw!
not as appealing in practicum... costly!