Imitation Learning

Why learn rewards?
- to optimise what the person wants
otherwise: just copy the human
  ✓ robot does thing in human like way
  ✓ robot does thing when we don't know R/V
  ✓ robot has model $\mathcal{G}$, the human policy

$z_d \sim \text{select from } \{ \mathbf{s}_d, \mathbf{a}_d, s'_d, a'_d \ldots \}$

Behavioral Cloning:
- train $\pi_1$ s.t. $\pi_1(s) = \pi_d(s)$
- parametrise $\pi_1$ by $\theta$, e.g. ANN

$\max_\theta \mathbb{E}_d \log \pi_\theta(a_d | s_d) \iff \max_\theta \sum_i \log \pi_\theta(a_i | s_i)$

$\min_\theta \mathbb{E} \left[ \text{KL}(\pi_d(\cdot|s) \parallel \pi_\theta(\cdot|s)) \right]$

$= \min_\theta \sum_i \frac{\pi_d(a_i | s_i)}{\pi_\theta(a_i | s_i)} \log \left( \frac{\pi_\theta(a_i | s_i)}{\pi_d(a_i | s_i)} \right)$

$\max_\theta \sum_i \log \pi_\theta(a_i | s_i)$
What's wrong w. BC?

- Assumes samples are iid, but we are in a sequential domain.
- Error accumulation

ARVIN CCMU):

Theoretical argument:

- Supervised learning has chance of E even if iid; T times: TE errors
- BC is not iid: errors accumulate → $T^2E$
  
  - Once error = cost of + @ each remaining timestep
  
  $E, CT - 1) + E, CT - 2) + \ldots + E, 1$
Fixes:
1. Don’t DAGGER (dataset aggregation)
   1. fit πθ on S
   2. roll it out, collect induced states S^π
   3. ask for action labels a_t^π
   4. S^π x a_t (S, π(St))

get “our policy” labels

turns out injecting noise in the demonstrator’s actions to get
them to go off and recover is probably enough (chapter 11?)

2) practice with R2 to get back on “we recover”.
2016 GAIL (generative adversarial imitation learning)

think of R2 as
max \min \left( \frac{\mathbb{E}[c(s,a)]}{\mathbb{I}} - \lambda + \gamma T \right)_C

want regularization \mathbb{E}[c] \rightarrow \mathbb{I} matches demand
state-action occupancy

idea: search for \mathbb{I} that does that more directly
$$DCS(a) = 0 \text{ if } (s,a) \text{ came from } \pi_D \text{, } \log DCS(a) = -\infty \text{ if } (s,a) \text{ came from } \pi_D \text{, }$$

else

chain $\tau$ to minimize, train $D$ to be close $\pi_D$ but high otherwise:

$$\max_{\tau} \min_D \left( \mathbb{E}_\tau \left[ \log DCS(a) \right] - \mathbb{E}_\pi \left[ \log (1 - DCS(a)) \right] \right)$$

iterate between gradient $D$ and $\tau_2$ update on $\pi_D$