learning from an expert:

- I have an MDP
- with UNKNOWN reward
- I see an expert acting.
- Q: learn policy

Versions

(1) Abbeel + Ng

assume $R(s) = \sum_i w_i \phi_i(s) = w \cdot \Phi(s)$

(ϕ_i: feature functions)

now any given policy π has a value

$$E_{s_0 \sim D}\left[V^{\pi}(s_0)\right] = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, \pi\right]$$

$$= w \cdot E\left[\sum_{t=0}^{\infty} \gamma^t \Phi(s_t) \,\middle|\, \pi\right]$$

So

$$\mu(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \Phi(s_t) \,\middle|\, \pi\right]$$

↑ determines value of policy

feature expectations.

When we have observations of an expert acting, we have state trajectories

we know

$$\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi(s_E^{(i)})$$

↑ estimates of feature expects for optimal policy.

strategy: search for policy π^* s.t.

$$\mu(\pi^*) \approx \hat{\mu}_E$$

Notice, this means

$$\max_{w;\|w\| \le 1} w^\top (\hat{\mu}_E - \mu(\pi^*)) \quad \text{small}$$

so search for π^*

s.t. $\max_{w;\|w\| \le 1} \left[w^\top (\hat{\mu}_E - \mu(\pi^*)) \right]$ is small

Search process:

- start w / π^0

- obtain

$$t^{(i)} = \max_{w;\ \|w\| \le 1} \left[\min_{j \in 0 \ldots i-1} \left\{ w \cdot \left(\hat{\mu}_E - \mu(\pi^j) \right) \right\} \right]$$

$w^{(i)}$ = the w that achieves this

w max difference between expert and best policy so far

- if $t^{(i)} < \varepsilon$

- now compute optimal policy for MDP using $R = (w^{(i)})^T \phi$

[known state transition probs — policy/value iteration]

[unknown — RL]

- estimate $\mu^{(i)}$

inner loop solves

$$\max_{t,w} \quad t$$

$$w^T \mu_E \geq w^T \mu^{(j)} + t$$

$$\|w\|_2 \leq 1$$

↑ quadratic program