Last Time

Sequential

This Time

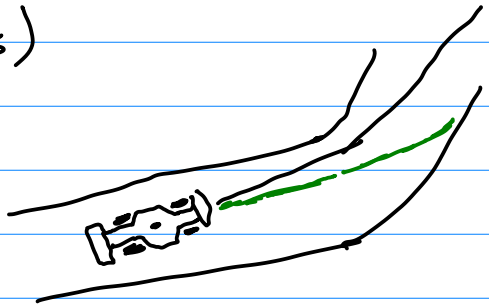
Imitation Learning
Inverse Reinforcement Learning

Behavioral Cloning

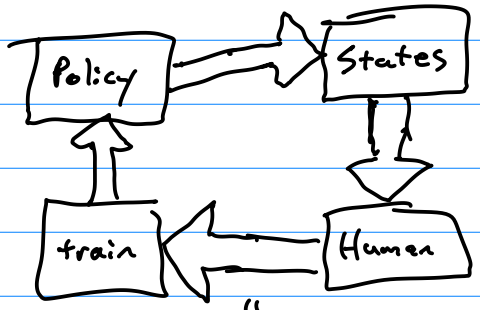
π_θ

$$\underset{\theta}{\text{maximize}} \prod_{(s,a) \in D} \pi_\theta(a|s)$$

Cascading Errors



Sequential Interactive Demonstration



Dagger "Dataset Aggregation"
SMILe

$\pi^{(1)} = \pi^E$

loop

execute $\pi^{(k)}$ to generate a dataset
query human response
train $\hat{\pi}^{(k)}$ with cloning
mix $\hat{\pi}^{(i)}$ with probability $\beta(1-\beta)^{(i-1)}$
to form $\hat{\pi}^{(k+1)}$

Apprenticeship Learning

Hard maneuver

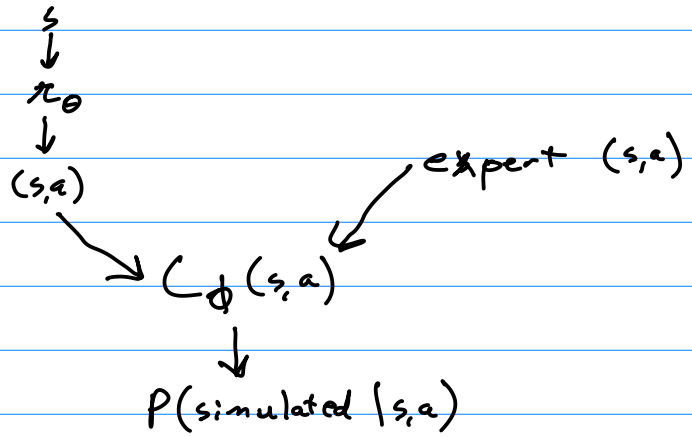
Follow with controller

Expert Trajectories

GAIL Generative Adversarial Imitation Learning

$\pi_\theta(a|s)$

$C_\phi(s,a)$ discriminator $s,a \rightarrow C \rightarrow P(\text{simulated from } \pi_\theta)$



$$\max_\phi \min_\theta \; \mathbb{E}_{(s,a)\sim\pi_\theta}\left[\log\left(C_\phi(s,a)\right)\right] + \mathbb{E}_{(s,a)\sim D}\left[\log\left(1 - C_\phi(s,a)\right)\right]$$

$(s,a)\sim D$: from expert

Inverse Reinforcement Learning

forward DMU

Given (S, A, R, T)

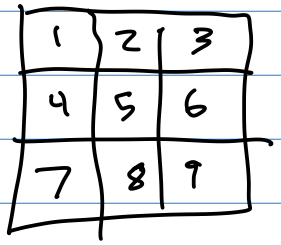
Find π^*

Inverse RL

Given $S, A, T, \{\tau\}_i$ ← samples from π^*

Find R

Breakout Rooms

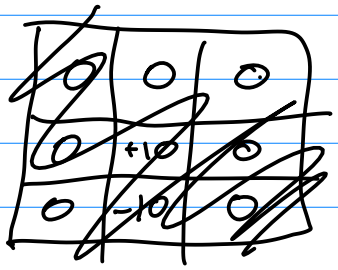


τ

1 → 2 → 3 ↓ 6 ↓ 9

1 → 2 ↓ 5 → 6 ↓ 9

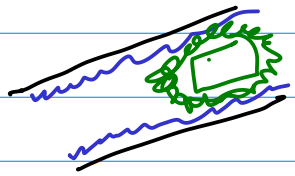
What is R?



0	-1	1
-5	1	1
-5	-5	+10

Under specified

$$R_\phi(s,a) = \phi^T \beta(s,a)$$



Maximum Margin IRL

β binary

$\|\phi\|_2 \leq 1$

$$E[U(\tau)] = E\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_\phi(s^{(k)}, a^{(k)})\right]$$

$$= E\left[\sum \gamma^{k-1} \phi^T \beta(s^{(k)}, a^{(k)})\right]$$

$$= \phi^T E\left[\sum \gamma^{(k-1)} \beta(s^{(k)}, a^{(k)})\right]$$

$$= \phi^T \mu_\pi$$

μ_π ← Feature Expectations

$\pi^{(1)}$ = random

while $t \geq \varepsilon$

$\mu^{(i)} \leftarrow$ feature expectations for $\pi^{(i)}$

$\phi^* \leftarrow \underset{\phi, t}{\text{maximize}}\ t$ ← margin

subject to $\phi^T \mu_E \geq \phi^T \mu^{(i)} + t$ for $i = 1, \ldots k-1$

$\|\phi\| \leq 1$

$\pi^{(i+1)} \leftarrow$ solve MDP with $R(s,a) = \phi^{*T} \beta(s,a)$

$\underset{\lambda}{\text{minimize}} \|\mu_E - \mu_\lambda\|_2 \qquad \mu_\lambda = \sum \lambda_i \mu^{(i)}$

subject to $\lambda \geq 0$

$\|\lambda\|_1 = 1$

Still Under-specified

Maximum Entropy IRL

Any policy introduces a distribution over trajectories $P_\pi(\tau)$

$$P_\phi(\tau) = \frac{1}{Z(\phi)} \exp\left(R_\phi(\tau)\right)$$

$$R_\phi(\tau) = \sum_t \gamma^{t-1} R_\phi(s^{(t)}, a^{(t)})$$

$$Z(\phi) = \sum_\tau \exp\left(R_\phi(\tau)\right)$$

normalization

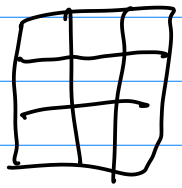
maximize $f(\phi)$

$$f(\phi) = \sum_{\tau \in D} \log P_\phi(\tau)$$

$$= \sum_{\tau \in D} \log \frac{1}{Z(\phi)} \exp\left(R_\phi(\tau)\right)$$

$$= \left(\sum_{\tau \in D} R_\phi(\tau)\right) - |D| \log Z(\phi)$$

optimize this with gradient ascent



Takeaways

Immitation Learning matches actions of an expert
- Behavioral Cloning
- Dagger
- GAIL

Inverse Reinforcement Learning
- Learning R based on Expert Trajectories
- X Underspecified
- Maximum Entropy