1) Inverse RL

like imititation learning, we learn from expert demonstrations.
rather than learning the expert's policy, IRL tries to learn the <u>reward function</u>

1) Imitation via reward fn
   - reward fn. is more succint/transferrable

2) Scientific Inquiry
   - modelling human/animal behavior

3) Multiagent setting - model other agents

<u>Setting</u>: (P known)

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, H, \mu\}$$

- Reward function $r(s,a)$ is unknown and signal r_t is unobserved
- observe trajectories from expert w/ optimal policy π^*

<u>Basic Idea</u>: Find a reward function which is consistent with the optimality of the expert policy

$$\text{find } r \quad \text{s.t.} \quad \mathbb{E}_{s,a\sim \underbrace{d_\mu^{\pi^*}}_{\text{estimate from expert trajectories}}}[r(s,a)] \geq \mathbb{E}_{s,a\sim d_\mu^{\pi}}[r(s,a)] \quad \underline{\forall \pi}$$

$r: \mathcal{S}\times\mathcal{A}\to[0,1]$

$\mathbb{E}_{s\sim\mu}[V^{\pi^*}(s)] =$

Problems w/ Formulation:

<u>1)</u> Need to consider all A^{SH} policies

2) Ambiguity: more than one reward function may satisfy $(r=0)$

Reframe: Find a policy that is as good as the expert

$$\underline{\text{Key assumption}}: \mathbb{E}[r(s,a)] = \overset{\text{unknown}}{\theta_*^T}\, \phi(s,a) \leftarrow \text{known}$$

linear reward wrt features

ex- $\phi(s,a) = \begin{bmatrix} \mathbb{P}(\text{building}) \\ \mathbb{P}(\text{sidewalk}) \\ \mathbb{P}(\text{road}) \\ \vdots \end{bmatrix}$ θ_* weighs negatives (driving on sidewalks hitting building) & positives (road)

We can write the policy consistency problem:

$$\text{find } \pi \quad \text{s.t.} \quad \underbrace{\mathbb{E}_{s,a \sim d^{\pi^*}_\mu}[\phi(s,a)]}_{\text{estimate from } \frac{1}{N}\sum_{i=1}^{N} \phi(s_i,a_i)} = \mathbb{E}_{s,a \sim d^{\pi}_\mu}[\phi(s,a)]$$

$\pi: \mathcal{S} \times \mathcal{A} \Rightarrow \Delta(\mathcal{S})$

To solve the ambiguity problem, we will use the "maximum entropy principle"

The Max Entropy IRL method:

$$\max_{\pi} \ \text{Entropy } \pi \leftarrow$$

$$\text{s.t.} \quad \mathbb{E}_{d^{\pi^*}_\mu}[\phi(s,a)] = \mathbb{E}_{d^{\pi}_\mu}[\phi(s,a)]$$

2) Maximum Entropy Principle

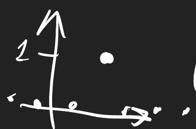
<u>Def</u> (Entropy) Distribution $P(x) \in \Delta(\mathcal{X})$

$$\text{Ent}(P) = \mathbb{E}_{x \sim P}[-\log(P(x))] = -\sum_{x \in \mathcal{X}} P(x)\log(P(x))]$$

$P(x) \in [0,1]$, Entropy is positive

ex - deterministic distribution $X = x_0$ w.p. 1

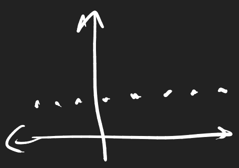
$$P_{x_0}(x) = \mathbb{1}\{x = x_0\}$$



$$\text{Ent}(P_{x_0}) = -\sum_{x \neq x_0} 0 \cdot \cancelto{0}{\log(0)} + -1 \cdot \cancelto{0}{\log(1)}$$

$$= 0$$

ex- uniform distribution over $|X| = N$ elements



$$\text{Ent}(U) = -\sum_{x \in X} \frac{1}{N} \log(1/N) = -\log(1/N)$$

$$= \log N$$

Max Ent Principle:

"Among consistent distributions, choose the one w/ the most uncertainty, i.e. the highest entropy.

w/ contraints arising from observation, mean, variance

The max-ent IRL approach:

$$\max_{\pi} \mathbb{E}_{s \sim d_M^\pi}[\text{Ent}(\pi(\cdot|s))] = \max_{\pi} -\mathbb{E}_{s \sim d_M^\pi}\left[\mathbb{E}_{a \sim \pi}[\log(\pi(a|s))]\right]$$

s.t. contraints

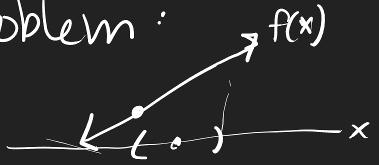
$$= \min_{\pi} \mathbb{E}_{s,a \sim d_M^\pi}[\log(\pi(a|s))]$$

$$\boxed{\begin{aligned} &\min_{\pi} \mathbb{E}_{s,a \sim d_M^\pi}[\log \pi(a|s)] \\ &\text{s.t. } \mathbb{E}_{s,a \sim d_M^{\pi^*}} \phi(s,a) = \mathbb{E}_{s,a \sim d_M^{\pi}} \phi(s,a) \end{aligned}}$$

$$\frac{1}{N}\sum_{i=1}^{N} \phi(s_i^*, a_i^*)$$

3) Constrained optimization

Consider the constrained optimization problem:



$$x^* = \underset{x}{\operatorname{argmin}} \left[f(x) \text{ s.t. } g(x) = 0 \right]$$

Lagrange Formulation:

$$\min_{x \in \mathbb{R}^d} \left[\max_{w \in \mathbb{R}} f(x) + w \cdot g(x) \right]$$

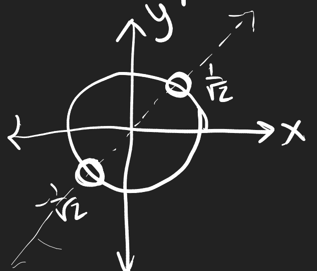
if $g(x) \neq 0$, $w \to \pm\infty$ so inner max is ∞

if $g(x) = 0$, inner maximitatin: $f(x)$

$$\max_{w} f(x) + w \cdot g(x) = \begin{cases} \infty & g(x) \neq 0 \\ f(x) & g(x) = 0 \end{cases}$$

$$x^* = \underset{x}{\operatorname{argmin}} \max_{w} f(x) + w \cdot g(x)$$

example: $\min x+y$ s.t. $x^2+y^2=1$



$$\Downarrow$$

$$\min_{x,y} \max_{w} \underbrace{x + y + w(x^2+y^2-1)}_{\mathcal{L}(x,y,w)}$$

$$\nabla_x \mathcal{L} = 1 + 2xw \overset{0}{\Rightarrow} x^* = \frac{-1}{2w_*}$$

$$\nabla_y \mathcal{L} = 1 + 2yw \overset{0}{\Rightarrow} y^* = \frac{-1}{2w_*}$$

$$\rightarrow \nabla_w \mathcal{L} = x^2 + y^2 - 1 \overset{0}{\Rightarrow} x^*)^2 + (y^*)^2 = 1$$

$$\left(\frac{-1}{2w_*}\right)^2 + \left(\frac{-1}{2w_*}\right)^2 = 1 \Rightarrow w_* = \pm\sqrt{1/2} = \frac{\sqrt{2}}{2}$$

critical points: $\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ and $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$

Iterative Procedure

```
initialize $w_0$
For $t = 0, \dots, T-1$
    $x_t = \arg\min_x f(x) + w_t g(x)$      (Best response)
    $w_{t+1} = w_t + \eta g(x_t)$      (incremental update)
Return $\bar{X} = \frac{1}{T} \sum_{t=0}^{T-1} x_t$,
```

$\bar{X} \to x^*$ as $T \to \infty$ if f, g are convex