CS 4/5789 25 Apr 2022 Prof Sarah Dean

Lecture 24: Inverse RL

1) Inverse RL Motivation & Setting

Like imitation learning, which we studied last week, the setting of inverse RL seeks to learn from expert demonstrations. However, rather than attempt to directly learn the expert's policy, IRL tries to learn the reward function. There may be several motivations for doing so:

1) scientific inquiry (modelling human or animal behavior)

our focus 2) Imitation learning via the reward function (reward function may be the most succinct & transferable information about a task, rather than a policy)

3) modelling other agents in a multiagent setting (adversarial or cooperative)

Setting: Finite Horizon MDP (P known)

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

Reward function $r(s,a)$ unknown / signal r_t unobserved
But observe traces of expert policy which is optimal π^*.

<u>Basic Idea</u>: Find reward functions which are consistent with expert being optimal:

$$\text{find } \underset{r: S\times\Omega \to [0,1]}{r} \quad \text{s.t.} \quad \underbrace{\mathbb{E}_{s,a\sim d_M^{\pi^*}}}_{\substack{\text{estimate from} \\ \text{expert trajectories}}}[r(s,a)] \geq \mathbb{E}_{s,a\sim d_M^{\pi}}[r(s,a)] \; \forall \pi.$$

Two Problems with this formulation:

1) As written, need to consider all possible policies (A^S)

2) Ambiguity: more than one reward function may satisfy this description (e.g. $r(s,a)=0$)

Since our goal is to ultimately use the learned reward function for policy design, we can reframe the consistency as a property of policies:

"find a policy that is at least as good as the expert."

<u>Key assumption</u>: $r(s,a) = \theta_*^\top \phi(s,a)$

The reward function is linear with respect to a known feature mapping

ex- $\phi(s,a) = \begin{bmatrix} \mathbb{P}(\text{building}) \\ \mathbb{P}(\text{sidewalk}) \\ \mathbb{P}(\text{road}) \\ \mathbb{P}(\text{person}) \end{bmatrix}$ and θ_* encodes that staying on the road is good (positive weight), sidewalk is bad (negative), colliding with person very bad (large negative)

Then we can write the policy consistency problem as.

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

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

By the linear cost assumption, this constraint implies that policies acheive the same reward.
(Exercise: write a short proof of this fact)

However, this does not solve the <u>ambiguity</u> problem: many such policies may satisfy this constraint.

Idea: among all policies satisfying the consistency constraint, choose the one that is the most "uncertain"

We will study the Maximum Entropy IRL method:

$$\max_{\pi} \ \text{Entropy of } \pi$$
$$\text{s.t.} \quad \mathbb{E}_{s,a \sim d^{\pi^*}_{\mathcal{M}}} [\varphi(s,a)] = \mathbb{E}_{s,a \sim d^{\pi}_{\mathcal{M}}} [\varphi(s,a)]$$

2) Maximum Entropy Priciple

our choice to choose the consistent policy with the maximum entropy follows from the "maximum entropy principle."

Definition (Entropy):

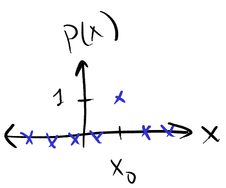
The entropy of a distribution $P \in \Delta(\mathcal{X})$ is defined as $\text{Ent}(P) = \mathbb{E}_{X \sim P}\left[-\log(P(X))\right] = -\sum_{x \in \mathcal{X}} P(x) \log(P(x))$

Because $P(x) \in [0,1]$, $\text{Ent}(P) \geq 0$.

Lower Entropy means lower uncertainty.

Ex: Deterministic distribution where $X = x_0$ w.p. 1.

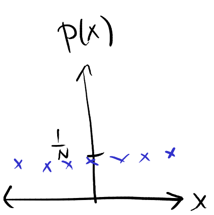
$$P_{x_0}(x) = \mathbb{1}\{x = x_0\}.$$



$$\text{Ent}(P_{x_0}) = -\sum_{x \in \mathcal{X}} \mathbb{1}\{x = x_0\} \log(\mathbb{1}\{x = x_0\})$$

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

Ex: uniform distribution over $|\mathcal{X}| = N$ elements.



$$U(x) = \frac{1}{N}$$

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

$$= N \cdot \frac{1}{N} \cdot \log(N) = \log(N)$$

Exercise: argue that the uniform distribution has the highest possible entropy.

Maximum entropy Priciple:

Among distributions consisent with constraints (i.e. obsewed data, mean, variance) choose the one with the most uncertainty, i.e. highest extropy.

This can be seen as an application of Occam's razor since we are in some sense making the fewest assumptions about the distribution.

ex - Gaussian distribution.

$$\max \operatorname{Ent}(P) \text{ s.t. } \mathbb{E}_{x \sim P}[x] = \mu \text{ and } \mathbb{E}[xx^T - \mu\mu^T] = \Sigma$$

The solution is a Gaussian Distribution (Proof / Derivation out of scope)

The max-ent IRL approach solves:

$$\max_{\pi} \mathbb{E}_{s \sim d^{\pi}_{\mu}}\left[\operatorname{Ent}(\pi(\cdot|s))\right] \longrightarrow \min_{\pi} \mathbb{E}\left[\log(\pi(a|s))\right]$$

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

we can simplify

$$\mathbb{E}_{s \sim d^{\pi}_{\mu}}\left[\operatorname{Ent}(\pi(\cdot|s))\right] = -\mathbb{E}_{s \sim d^{\pi}_{\mu}}\left[\mathbb{E}_{a \sim \pi(\cdot|s)}\left[\log(\pi(a|s))\right]\right] = -\mathbb{E}_{s,a \sim d^{\pi}_{\mu}}\left[\log(\pi(a|s))\right]$$

3) Constrained Optimization

We've considered many optimization algorithms throughout the semester but not many with constraints (exception: trust regions)

Consider the constrained optimization problem:

$$x^* = \arg\min_x f(x) \quad \text{s.t.} \quad g(x) = 0 \qquad \text{(primal)}$$

To solve this problem we consider the Lagrange formulation, which converts it into an unconstrained:

$$\min_x \left[\max_w f(x) + w\,g(x) \right] \qquad \text{(Lagrange)}$$

Now if $g(x) \neq 0$ (i.e. x is infeasible for primal), then

$$\max_w f(x) + w \cdot g(x) = \infty \quad (w^* = \infty).$$

And if $g(x) = 0$ (i.e. x is feasible) then

$$\max_w f(x) + w\,g(x) = f(x).$$

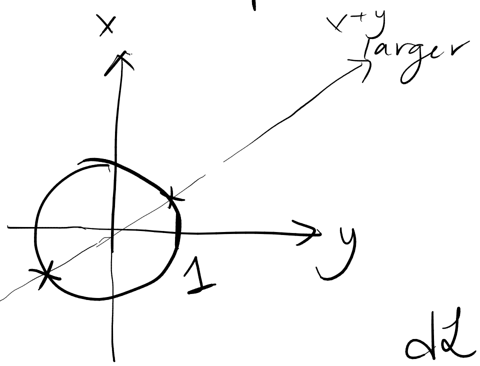
Then the Lagrange formulation encodes

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

Since ∞ is undesirable with respect to the outermost minimization, solving the Lagrange formulation is equivalent!

$$x^* = \arg\min_x \left[\max_w f(x) + w \cdot g(x) \right]$$

Example:



$$\min\ x+y \quad \text{s.t.} \quad x^2+y^2=1$$

$$\mathcal{L}(x,y,w) = x+y+w(x^2+y^2-1)$$

$$\min_{x,y}\ \max_{w}\ \mathcal{L}(x,y,w)$$

$$\frac{\partial \mathcal{L}}{\partial x} = 1+2wx \overset{=0}{\Rightarrow} x^* = \frac{-1}{2w}$$

$$\frac{\partial \mathcal{L}}{\partial y} = 1+2wy \overset{=0}{\Rightarrow} y^* = \frac{-1}{2w}$$

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

solving system of three equations:

$$1/4w^2 + 1/4w^2 = 1 \iff w = \pm\sqrt{1/2}$$

The critical points are

$$(-1/\sqrt{2}, -1/\sqrt{2}) \qquad (1/\sqrt{2}, 1/\sqrt{2})$$

minimum — maximum

Iterative Procedure:

For $t=0,\dots,T-1$:

$$x^t = \arg\min_x\ f(x) + w^t g(x) \quad \text{(Best response)}$$

$$w^{t+1} = w^t + \eta g(x) \quad \text{(Incremental update)}$$

Return $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$

Informal Theorem: $\bar{x} \to x^*$ as $T \to \infty$ if f,g convex