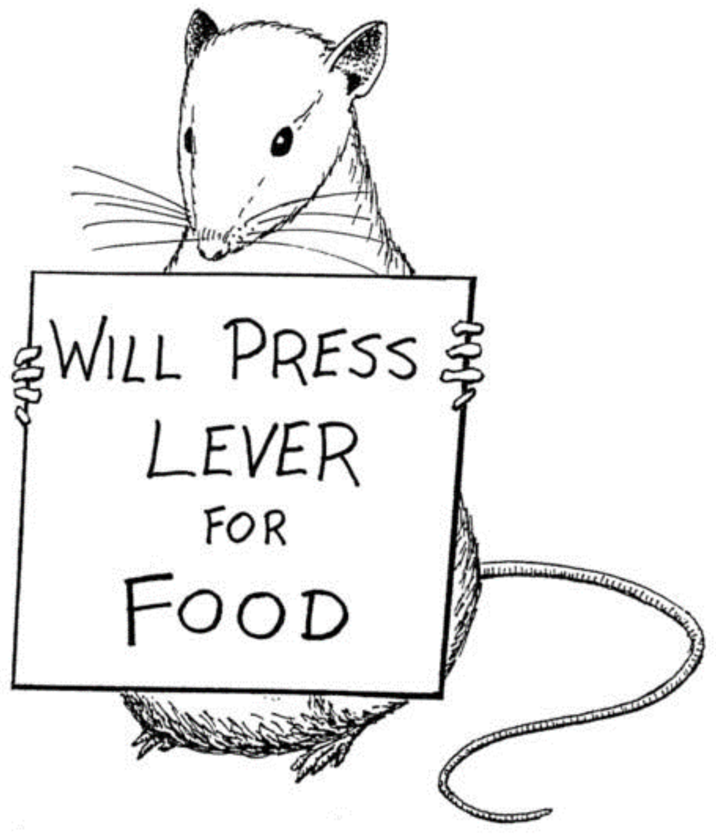
University of California, Irvine

CS 277: Control and Reinforcement Learning

Winter 2021

Lecture 14: Inverse RL

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine



Today's lecture

Feature Matching

MaxEnt IRL

GAIL

Learning rewards from demonstrations

- RL: rewards → policy; IL: demonstrations → policy
- Inverse Reinforcement Learning (IRL): demonstrations → reward function
  - Better understand agents (humans, animals, users, markets)
    - Preference elicitation, teleology (the "what for" of actions), theory of mind, language
  - First step toward Apprenticeship Learning: demos → rewards → policy
    - Infer the teacher's goals and learn to achieve them; overcome suboptimal demos
    - Partly model-based (learn r but not p); may be easier to learn, generalize, transfer
    - Teacher and learner can have different action spaces (e.g., human → robot)

Inverse Reinforcement Learning (IRL)

- Given a dataset of demonstration trajectories $\mathcal{D} = \{\xi_i\}$

$r(s)$ **expressive enough**

- Find teacher's reward function $r : \mathcal{S} \times \cancel{\mathcal{A}} \to \mathbb{R}$
    - Principle: demonstrated actions should achieve high expected return
- IRL is ill-defined
    - How low is the reward for states and actions not in $\mathcal{D}$?
    - How is the reward distributed along the trajectory?
        - Sparse rewards = identify "subgoal" states; dense = score each step, as hard as IL
    - Demonstrator can be fallible = take suboptimal actions; how much?

Feature matching

- Assume linear reward $r_\theta(s) = \theta^\intercal f_s$ in oracle state features $f_s \in \mathbb{R}^d$

  $t \sim \mathrm{Geom}(1-\gamma)$

  up to $\cdot (1-\gamma)$

  - $\implies$ Return = $R_{\pi;\theta} = \sum_t \gamma^t \mathbb{E}_{s_t \sim p_\pi}[\theta^\intercal f_{s_t}] = \mathbb{E}_{s \sim p_\pi}[\theta^\intercal f_s]$ (with $p_\theta(s) = \sum_t \gamma^t p_\theta(s_t)$)

- Teacher optimality: return $R_{e;\theta}$ higher than any other policy's return $R_{\pi;\theta}$
  - $\implies$ Find θ that maximizes the gap $R_{e;\theta} - R_{\pi;\theta}$ (for which π?)
  - $\implies$ Apprenticeship Learning: find π that maximizes $R_{\pi;\theta}$ (for which θ?)

- Solve: $\max_\theta \min_\pi \{R_{e;\theta} - R_{\pi;\theta}\} = \max_\theta \min_\pi \{\mathbb{E}_{s \sim p_e}[\theta^\intercal f_s] - \mathbb{E}_{s \sim p_\pi}[\theta^\intercal f_s]\}$
  - Approximate $s \sim p_e$ with $s \sim \mathscr{D}$

Feature matching

- Feature Matching:
  - Initialize $\Pi = \{\pi_0\}$
  - Repeat:
    - Solve the Quadratic Program: $\max_{\eta, \|\theta\|_2 \leq 1} \eta$ s.t. $\mathbb{E}_{s\sim\mathcal{D}}[\theta^\intercal f_s] \geq \mathbb{E}_{s\sim p_\pi}[\theta^\intercal f_s] + \eta \quad \forall \pi \in \Pi$
    - Add to Π the optimal policy π for $r_\theta(s) = \theta^\intercal f_s$
- On convergence: π optimal for θ (no gap), can't find θ with gap
  - $\implies \mathbb{E}_{s\sim\mathcal{D}}[\theta^\intercal f_s] \approx \mathbb{E}_{s\sim p_\pi}[\theta^\intercal f_s]$ for all $\theta \implies \mathbb{E}_{s\sim\mathcal{D}}[f_s] \approx \mathbb{E}_{s\sim p_\pi}[f_s]$ **feature matching**

Today's lecture

Feature Matching

MaxEnt IRL

GAIL

Modeling bounded teachers

- An expert teacher maximizes the return $R_{e;\theta} = \sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim p_e}[\theta^\intercal f_{s_t}] = \mathbb{E}_{\xi \sim p_e}[\theta^\intercal f_\xi]$
  - With the trajectory-summed features $f_\xi = \sum_t f_{s_t}$
- Bounded rationality: teacher has "unintentional" prior policy π_0
  - Cost to intentionally diverge: $\mathbb{D}[\pi_e \| \pi_0]$ (with π_0 uniform: $\mathbb{H}[\pi_e]$)
  - Total cost over trajectory: $\mathbb{D}[p_e(\xi) \| p_0(\xi)] = \mathbb{E}_{\xi \sim p_e}\left[\log \frac{p_e(\xi)}{p_0(\xi)}\right] = \sum_t \underbrace{\mathbb{E}_{s_t \sim p_e}\left[\log \frac{\pi_e(a_t | s_t)}{\pi_0(a_t | s_t)}\right]}_{\mathbb{D}[\pi_e \| \pi_0]}$
- Bounded optimality: $\max_{\pi_e} \mathbb{E}_{\xi \sim p_e}[\theta^\intercal f_\xi] - \tau \mathbb{D}[p_e \| p_0]$

Bounded optimality: naïve solution

- Bounded optimality: $\max_{\cancel{\pi_e}\ p_e} \mathbb{E}_{\xi \sim p_e}[\theta^\intercal f_\xi] - \mathbb{D}[p_e \| p_0]$
  - Naïve solution: allow any distribution p_e over trajectories
  - No need to be consistent with dynamics $p(s' | s, a) \implies p_e$ may be unachievable
- Add the constraint $\sum_\xi p_e(\xi) = 1$ with Lagrange multiplier λ
- Differentiate by $p_e(\xi)$ and $= 0$ to optimize

$$\theta^\intercal f_\xi - \log p_e(\xi) + \log p_0(\xi) - 1 + \lambda = 0 \implies p_e(\xi) = \frac{p_0(\xi)\exp(\theta^\intercal f_\xi)}{\sum_{\bar{\xi}} p_0(\bar{\xi})\exp(\theta^\intercal f_{\bar{\xi}})}$$

IRL with bounded teacher

- Assume that demonstrations are distributed $p_\theta(\xi) = \frac{1}{Z_\theta} p_0(\xi) \exp(\theta^\intercal f_\xi)$
  - With partition function $Z_\theta = \mathbb{E}_{\xi \sim p_0}[\exp(\theta^\intercal f_\xi)]$
- Find θ that minimizes NLL of demonstrations

$$\nabla_\theta \log p_\theta(\xi) = \nabla_\theta(\theta^\intercal f_\xi - \log Z_\theta) = f_\xi - \frac{1}{Z_\theta} \nabla_\theta Z_\theta$$

$$= f_\xi - \frac{1}{Z_\theta} \mathbb{E}_{\bar{\xi} \sim p_0}[\exp(\theta^\intercal f_{\bar{\xi}}) f_{\bar{\xi}}] = f_\xi - \mathbb{E}_{\bar{\xi} \sim p_\theta}[f_{\bar{\xi}}]$$

  - To compute gradient, we need $p_\theta \implies$ we need Z_θ

Computing Z_θ: backward recursion

- Partition function: $Z_\theta = \mathbb{E}_{\xi\sim p_0}[\exp(\theta^\intercal f_\xi)]$

- Compute Z_θ recursively backward:

$$Z_\theta(s_t, a_t) = \mathbb{E}_{p_0}[\exp(\theta^\intercal f_{\xi \geq t}) \,|\, s_t, a_t]$$

$$Z_\theta(s_t) = \mathbb{E}_{p_0}[\exp(\theta^\intercal f_{\xi \geq t}) \,|\, s_t]$$

- Z_θ defines $p_\theta(\xi) = \frac{1}{Z_\theta} p_0(\xi) \exp(\theta^\intercal f_\xi)$

  - Marginalizing: $\pi_\theta(a_t \,|\, s_t) = \pi_0(a_t \,|\, s_t) \frac{Z_\theta(s_t, a_t)}{Z_\theta(s_t)}$

- π_θ is not globally consistent $p_\theta(\xi) \neq p_{\pi_\theta}(\xi)$, because we ignored the dynamics

Computing Z_θ: backward recursion

- Partition function: $Z_\theta = \mathbb{E}_{\xi \sim p_0}[\exp(\theta^\intercal f_\xi)]$

- Compute Z_θ recursively backward:

$$Z_\theta(s_t, a_t) = \mathbb{E}_{p_0}[\exp(\theta^\intercal f_{\xi \geq t}) \mid s_t, a_t] = \exp(\theta^\intercal f_{s_t}) \mathbb{E}_{s_{t+1} \mid s_t, a_t \sim p}[Z_\theta(s_{t+1})]$$

$$Z_\theta(s_t) = \mathbb{E}_{p_0}[\exp(\theta^\intercal f_{\xi \geq t}) \mid s_t] = \mathbb{E}_{a_t \mid s_t \sim \pi_0}[Z_\theta(s_t, a_t)]$$

- Z_θ defines $p_\theta(\xi) = \frac{1}{Z_\theta} p_0(\xi) \exp(\theta^\intercal f_\xi)$

  - Marginalizing: $\pi_\theta(a_t \mid s_t) = \pi_0(a_t \mid s_t) \frac{Z_\theta(s_t, a_t)}{Z_\theta(s_t)}$

- π_θ is not globally consistent $p_\theta(\xi) \neq p_{\pi_\theta}(\xi)$, because we ignored the dynamics

MaxEnt IRL

- For each sample $\xi \sim \mathscr{D}$:
  - Compute $Z_\theta = \mathbb{E}_{\xi \sim p_0}[\exp(\theta^\intercal f_\xi)]$ recursively backward
  - Compute $\mathbb{E}_{\bar{\xi} \sim p_{\pi_\theta}}[f_{\bar{\xi}}]$ recursively forward
  - Take a gradient step to improve θ: $\nabla_\theta \log p_\theta(\xi) \approx f_\xi - \mathbb{E}_{\bar{\xi} \sim p_{\pi_\theta}}[f_{\bar{\xi}}]$
- At the optimum: feature matching $\mathbb{E}_{\xi \sim \mathscr{D}}[f_\xi] = \mathbb{E}_{\xi \sim p_{\pi_\theta}}[f_\xi]$
  - MaxEnt IRL approximates $\max_\theta \mathbb{H}[\pi_\theta]$ s.t. $\mathbb{E}_{\xi \sim \mathscr{D}}[f_\xi] = \mathbb{E}_{\xi \sim p_{\pi_\theta}}[f_\xi]$

Limitations:

- **Requires dynamics** p
- **Assumes** $p_\theta = p_{\pi_\theta}$
- **Assumes** $\mathscr{D} = p_e$

Today's lecture

Feature Matching

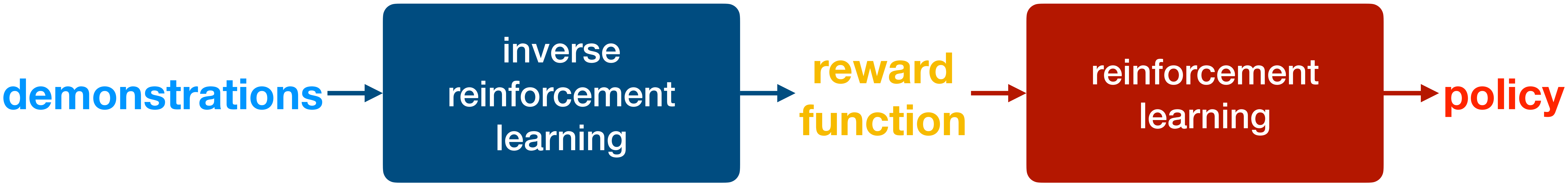
MaxEnt IRL

GAIL

IRL: downstream tasks

- Motivation: learn reward function for downstream tasks...

...such as RL



- IL = RL ∘ IRL (composition of RL on IRL)
- Our algorithms already learn π as part of learning θ for $r : s \mapsto \theta^\intercal f_s$
  - Let's directly optimize IRL for the overall IL task = learn good π

IL as RL ∘ IRL

- Entropy-regularized RL: $\max_{\pi \in \Pi} \left\{ \mathbb{E}_{s \sim p_\pi}[r(s)] + \mathbb{H}[\pi] \right\}$

- MaxEnt IRL: $\max_{r \in \mathbb{R}^{\mathcal{S}}} \left\{ \mathbb{E}_{s \sim p_e}[r(s)] - \max_{\pi \in \Pi} \left\{ \mathbb{E}_{s \sim p_\pi}[r(s)] + \mathbb{H}[\pi] \right\} \right\} - \psi(r)$

regularization over reward function space

- For any π, our objective with respect to r is:

$$\psi^*(p_e - p_\pi) = \max_{r \in \mathbb{R}^{\mathcal{S}}} \left\{ \overbrace{(p_e - p_\pi)}^{\in \mathbb{R}^{\mathcal{S}}} \cdot r - \psi(r) \right\}$$

  - This form of function $\psi^* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}$ is called the convex conjugate of ψ

Reward-function regularizers

$$\psi^*(p_e - p_\pi) = \max_{r \in \mathbb{R}^{\mathcal{S}}} \left\{ (p_e - p_\pi) \cdot r - \psi(r) \right\}$$

- Without regularizer: $\psi = 0 \implies$ solution only exists when $p_e = p_\pi$
  - $\implies$ learner achieves teacher's state distribution: perfect solution, but hard to find
- Hard linearity constraint: $\psi(r) = \begin{cases} 0 & \text{if } r(s) = \theta^\intercal f_s \\ \infty & \text{otherwise} \end{cases}$
  - $\implies$ max-entropy feature matching (MaxEnt IRL)
  - Great when the reward function really is linear in f_s, otherwise no guarantees

Generative Adversarial Networks (GANs)

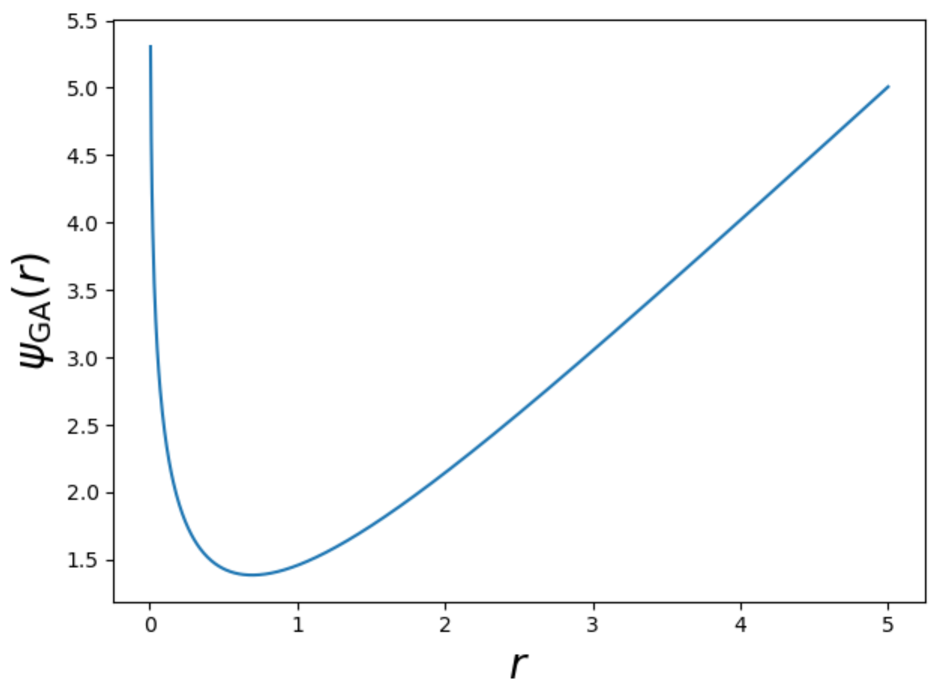
- Train generative model $p_\theta(s)$ to generate states / observations
  - Can we focus the training on failure modes?

- Also train discriminator $D_\phi(s) \in [0,1]$ to score instances
  - Kind of like a critic: are generated instances good?
- $D_\phi(s)$ predicts the probability $p(s \text{ generated by learner} \mid s) = \dfrac{p_\theta(s)}{p_\theta(s) + p_e(s)}$
  - Trained with cross-entropy loss: $\max_\phi \left\{ \mathbb{E}_{s \sim p_\theta}[\log D_\phi(s)] + \mathbb{E}_{s \sim p_e}[\log(1 - D_\phi(s))] \right\}$
- The generator tries to fool the discriminator: $\min_\theta \mathbb{E}_{s \sim p_\theta}[\log D_\phi(s)]$

Teacher-based reward-function regularizer

- Consider the regularizer

$$\psi_{\mathrm{GA}}(r) = \mathbb{E}_{s\sim p_e}[r(s) - \log(1 - \underbrace{\exp(-r(s))}_{D(s)})]$$



- It's convex conjugate is:

$$\psi^*_{\mathrm{GA}}(p_e - p_\pi) = \max_{r\in\mathbb{R}^{\mathcal{S}}} \left\{(p_e - p_\pi)\cdot r - \psi(r)\right\}$$

$$= \max_{r\in\mathbb{R}^{\mathcal{S}}}[r(s) - r(s) + \log(1 - D(s))] - \mathbb{E}_{s\sim p_\pi}[\overbrace{r(s)}^{-\log D(s)}]$$

$$= \mathbb{E}_{s\sim p_\pi}[\log D(s)] + \mathbb{E}_{s\sim p_e}[\log(1 - D(s))]$$

  - $\implies$ GAN: generator p_π imitating teacher p_e; discriminator $D(s) = \exp(-r(s))$

Generative Adversarial Imitation Learning (GAIL)

Input: demonstration dataset $\mathcal{D}_T \sim p_T$

repeat

$\quad \mathcal{D}_L \leftarrow$ roll out π_θ

$\quad$ take discriminator gradient ascent step

$$\mathbb{E}_{s\sim\mathcal{D}_L}[\nabla_\phi \log D_\phi(s)] + \mathbb{E}_{s\sim\mathcal{D}_T}[\nabla_\phi \log(1 - D_\phi(s))]$$

$\quad$ take entropy-regularized policy gradient step with reward $r(s) = -\log D_\phi(s)$

- We've already seen one entropy-regularized PG algorithm: TRPO
  - More next time

Recap

- To understand behavior: infer the intentions of observed agents
- If teacher is optimal for a reward function
  - The reward function should make an optimizer imitate the teacher
  - State (or state–action) distribution of learner should match the teacher
- In this view, Inverse Reinforcement Learning (IRL) is a game:
  - Reward is optimized to show how much the teacher is better than the learner
  - Learner optimizes for the reward
  - Reward is like a discriminator (high = probably teacher); learner like a generator