Last Time

- Multi Obj
  - Constrained
  - Linear Weighted
- Risk Aware
- Pure Info Gathering

This time: VOI, IRL

o already, o' value of observing

$$VOI(o' \mid o) = \sum_{o'} \left(P(o' \mid o) EU^*(o, o') \right) - EU^*(o)$$

Example

$a \sim U([-\pi, \pi])$

$$U(c) = \begin{cases} 1 & \text{if } |a - c| \leq \pi/4 \\ 0 & \text{o.w.} \end{cases}$$

VOI(A)?

$$VOI(A) = \int_a P(a) EU^*(a)\, da - EU^*()$$

$P(a) = \frac{1}{2\pi}$, $EU^*(a) = 1$, $EU^*() = \frac{1}{4}$

$$= 1 - \frac{1}{4}$$

$$= \frac{3}{4}$$

$$VOI(o' \mid b) = \sum_{o'} P(o' \mid b) V^*(\tau(b, o)) - V^*(b)$$

Inverse RL IRL in IRL

forward RL

given S, A, T, R

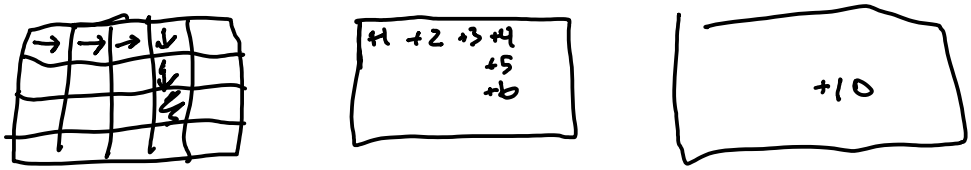
find π^*

inverse

given $S, A, T, \{\tau_i\}$ ← trajectories sampled with π^*

find R

Imitation Learning Reinforcement Learning



Underspecification

Linear $R_\psi(s,a) = \sum_i \psi_i f(s,a_i)$

s — 0 0 0
a — 0 0 0
0 0 0

$R_\psi(s,a)$

↖ parameters ψ

---

4 approaches

Feature matching IRL

π^{r_ψ} = optimal policy for r_ψ

pick ψ s.t. $E_{\pi^{r_\psi}}[f(s,a)] = E_{\pi^*}[f(s,a)]$

↖ state-action marginal distribution ↖ from samples

still ambiguous

Maximum margin

maximize m
ψ, m

subject to $\psi^T E_{\pi^*}[f(s,a)] \geq \max_{\pi} \psi^T E_{\pi}[f(s,a)] + m$

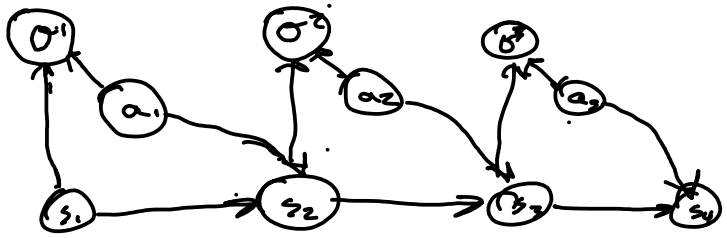
$\Leftrightarrow$

minimize $\frac{1}{2}\|\psi\|^2$
ψ

s.t $\psi^T E_{\pi^*}[f(s,a)] \geq \max_{\pi} \psi^T E_{\pi}[f(s,a)] + D(\pi, \pi^*)$

- kind of arbitrary
- No model of expert subopt.
- Messy Constrained Opt.

1.



$p(O_t) \propto \exp(r(s_t, a_t))$

$\pi(a_t | s_t) \propto e^{Q_t(s_t,a_t) - V_t(s_t)} = e^{A_t(s_t,a_t)}$

$$p(\tau, O_{1:T}) = \frac{p(\tau, O_{1:T})}{p(O_{1:T})}$$

$$\propto p(\tau) \prod_t \exp(r(s_t, a_t)) = p(\tau) \exp\left(\sum_t r(s_t, a_t)\right)$$

$$p(\tau | O_{1:T}, \psi) \propto p(\tau) \exp\left(\sum_t r_\psi(s_t, a_t)\right)$$

$$\max_\psi \frac{1}{N} \sum_{i=1}^{N} \log p(\tau_i | O_{1:T}, \psi)$$

$$= \max_\psi \frac{1}{N} \sum_i r_\psi(\tau_i) - \log Z$$

Z ← normalization "partition function" hard part

$$Z = \int p(\tau) \exp(r_\psi(\tau)) d\tau$$

$$\nabla_\psi \mathcal{L} = \frac{1}{N} \sum \nabla_\psi r_\psi(\tau_i) - \frac{1}{Z} \int p(\tau) \exp(r_\psi(\tau)) \nabla_\psi r_\psi(\tau) d\tau$$

$$= \mathop{E}_{\tau \sim \pi^*(\tau)} [\nabla_\psi r_\psi(\tau_i)] - \mathop{E}_{\tau \sim p(\tau | \mathcal{O}_{1:T}, \psi)} [\nabla_\psi r_\psi(\tau)]$$

↑ soft optimal policy for current reward.

2.

$$\beta_t(s_t, a_t) = p(\mathcal{O}_{1:T} | s_t, a_t)$$ "backward messages" can calculate with DP

$$\alpha_t(s_t) = p(s_t | \mathcal{O}_{1:t-1})$$ "forward messages"

Max Ent RL

3.

1. Given ψ, compute $\beta(s_t, a_t)$
2. Given ψ, β, compute $\alpha(s_t)$
3. $\mu_t(s_t, a_t) \propto \beta(s_t, a_t)\, \alpha(s_t)$
4. Eval $\nabla_\psi \mathcal{L} \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\psi r_\psi(s_{i,t}, a_{i,t}) - \sum_t \int\int \mu_t(s_t, a_t) \nabla_\psi r_\psi(s_t, a_t)\, ds_t\, da_t$
5. $\psi \leftarrow \psi + \eta \nabla_\psi \mathcal{L}$

Why "Max Ent"?

if $r_\psi(s_t, a_t) = \psi^T f(s_t, a_t)$

maximizes ψ $\mathcal{H}(\pi^{r_\psi})$ s.t. $E_{\pi^{r_\psi}}[f] = E_{\pi^*}[f]$

3.

3. Guided Cost Learning

Problem with Max Ent
1. Solving for soft policy in inner loop
2. Known dynamics

What if we can only sample?

$$\nabla_\psi \mathcal{L} = \underset{\tau \sim \pi^*}{E}[\nabla_\psi r_\psi(\tau_i)] - \underset{\tau \sim p(\tau|\sigma_{1:T},\psi)}{E}[\nabla_\psi r_\psi(\tau)]$$

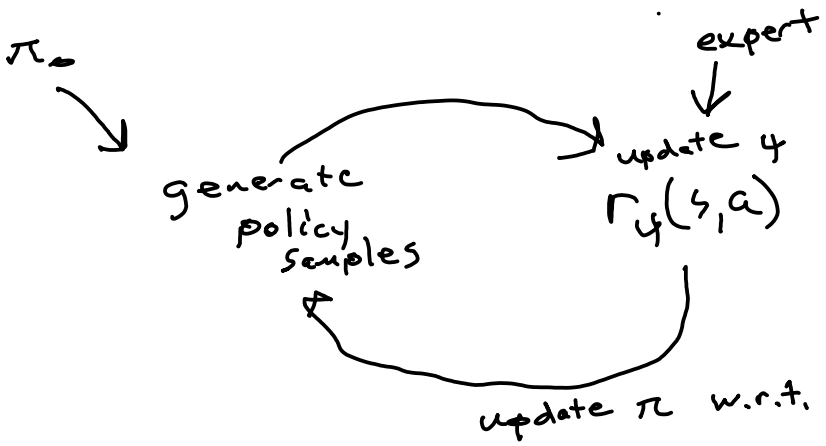
Solve with RL ≈ Do 1 step of RL

$$\approx -\frac{1}{M}\sum_{j=1}^{M} \nabla_\psi r_\psi(\tau_i)$$

$$\frac{1}{\sum w_j}\sum_j w_j \nabla_\psi r_\psi(\tau_i)$$

$$w_j = \frac{p(\tau)\exp(r_\psi(\tau_j))}{\pi(\tau_i)}$$

Guided Cost Learning



4 GAN

Looks like a game

policy tries to make it harder to distinguish from Demos

reward learning tries to make human policies more likely

Generative Adversial Network

Generator Discriminator

policy reward learner .