Inverse Reinforcement Learning
(Inverse Optimal Control)

TODO
↳ put in MDP lingo as well

optimal control: given $\mathcal{U}$, find $\xi^* = \arg\min \mathcal{U}[\xi]$

IOC: given ξ_D, find $\mathcal{U}: \Xi \to \mathbb{R}^+$ s.t. $\mathcal{U}[\xi_D] \leq \mathcal{U}[\xi], \forall \xi \in \Xi_{s \to g}$

find a cost function(al) that explains the demonstration

why? e.g. $g \to \hat{g}$ $\quad \hat{\xi} = \underset{\xi \in \Xi_{s \to \hat{g}}}{\arg\min} \mathcal{U}[\xi]$

→ so that we can generalize to new problems

example applications: - driving
- predictable motion
- anything where it's harder to write down tradeoffs than demonstrate behavior

MMP - Maximum Margin Planning

$\mathcal{U}[\xi_D] \leq \mathcal{U}[\xi] \;\forall \xi \in \Xi_{s \to g}$

$\mathcal{U}[\xi_D] \leq \min_{\xi} \mathcal{U}[\xi] \quad (1)$

Problem: $\mathcal{U}[\xi] = k, \forall \xi$ solves (1)

Solution: make $\mathcal{U}[\xi_D]$ better by a margin

$\mathcal{U}[\xi_D] \leq \min_{\xi} \left[\mathcal{U}[\xi] - \ell[\xi, \xi_D] \right]$

with ℓ smaller when ξ closer to ξ_D

↳ give everything but ξ_D an advantage

e.g. $\ell[\xi, \xi_D] = \begin{cases} 0, & \text{if } \xi = \xi_D \\ 1, & \text{otherwise} \end{cases}$

in practice: ℓ smooth, eg L_2 norm

$$\max_{\mathcal{U}} \min_{\xi} [\mathcal{U}[\xi] - \ell[\xi, \xi_D]] - \mathcal{U}[\xi_D]$$

$$\min_{\mathcal{U}} \mathcal{U}[\xi_D] - \min_{\xi} [\mathcal{U}[\xi] - \ell[\xi, \xi_D]] + \frac{\lambda}{2} \text{Regularization}(\mathcal{U})$$

Parametrize $\mathcal{U}$, search its parameters

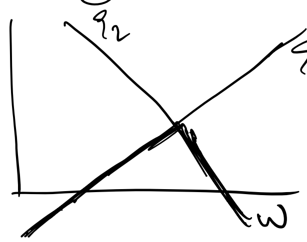
$\mathcal{U}[\xi] = w^T f_\xi$ f_ξ feature vector, e.g. $\begin{bmatrix} \text{length} \\ \text{obs dist} \\ \text{terrain} \end{bmatrix}$

$$\min_w \underbrace{w^T f_{\xi_D} - \min_{\xi} [w^T f_\xi - \ell(\xi, \xi_D)] + \frac{\lambda}{2} \|w\|^2}_{c(w)}$$

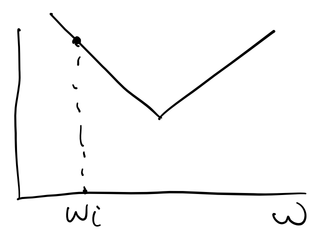
$\min_w c(w)$ $w_{i+1} = w_i - \alpha \nabla_{w_i} c$

$-\min_{\xi} [w^T f_\xi - \ell[\xi, \xi_D]]$ piecewise linear, convex:

$w^T f_\xi - \ell_\xi$



$-\min [\ldots]$



Let $\xi^*_w = \arg\min_{\xi} [w^T f_\xi - \ell(\xi, \xi_D)]$ $w_i \to \xi^*_{w_i}$

$$\nabla_{w_i} (-\min_{\xi} [w^T f_\xi - \ell[\xi, \xi_D]] = \nabla_{w_i} (-(w^T f_{\xi^*_{w_i}} - \ell[\xi^*_w, \xi_D]))$$

"subgradient"

$$\nabla_{w_i} c = f_{\xi_D} - f_{\xi^*_{w_i}} + \lambda w_i$$

$$w_{i+1} = w_i - \underbrace{\alpha \lambda w_i}_{\text{shrink } w} - \alpha \underbrace{(f_{\xi_D} - f_{\xi^*_{w_i}})}_{\text{gets you to match } f_{\xi_D}}$$

example: ξ_D goes on grass $\phi_{\xi_D} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{matrix} \text{grass} \\ \text{rock} \end{matrix}$ $\Big\} \Rightarrow$

$\xi^*_{w_i}$ goes on rock $\phi_{\xi^*_{w_i}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$\Rightarrow w_{i+1} = w_i(1-2\lambda) - \alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

$w(0)$ - grass - <u>decreases</u>
$w(1)$ - rock - <u>increases</u> $\Big\} \Rightarrow$

$\Rightarrow$ grass becomes <u>cheaper</u>; rock becomes more expensive

Maximum Entropy IRL

- assume H is not optimal, and nothing else:

entropy

$$\max_P H(P)$$

$$\text{s.t. } \mathbb{E}_{\xi_D \sim P}[u[\xi_D]] = u^* + \varepsilon$$

rationality coeff (dep. on ε)

$$\hookrightarrow P(\xi \mid w) = \frac{e^{-\beta w^T \phi_\xi}}{\sum_{\bar\xi} e^{-\beta w^T \phi_{\bar\xi}}}$$

partition function

$\beta = 0$ — $\beta \to \infty$

uniform human

- ξ_D is not perfect :

$$P(\xi \mid \omega) \propto e^{-\omega^T f_\xi}$$

- MLE :

$$\max_{\omega} P(\xi_D \mid \omega)$$

$$\Leftrightarrow \max_{\omega} \log P(\xi_D \mid \omega)$$

$$\Leftrightarrow \max_{\omega} \log \frac{e^{-\omega^T f_{\xi_D}}}{\sum_{\xi} e^{-\omega^T f_\xi}}$$

$$\Leftrightarrow \max_{\omega} -\omega^T f_{\xi_D} - \log \sum_{\xi} e^{-\omega^T f_\xi}$$

$$\nabla_\omega : -f_{\xi_D} - \frac{1}{\sum_{\xi} e^{-\omega^T f_\xi}} \sum_{\xi} \left[e^{-\omega^T f_\xi} \cdot (-f_\xi) \right]$$

$$\nabla_\omega : -f_{\xi_D} - \sum_{\xi} \frac{e^{-\omega^T f_\xi}}{\sum_{\bar{\xi}} e^{-\omega^T f_{\bar{\xi}}}} (-f_\xi)$$

$$\nabla_\omega : -\left(f_{\xi_D} - \sum_{\xi} P(\xi \mid \omega) f_\xi \right)$$

$$\underbrace{\omega_{t+1} = \omega_t -}_{\text{ascent?}} \alpha \left(f_{\xi_D} - \underbrace{\mathbb{E}_{\xi \sim P(\xi \mid \omega_t)} f_\xi}_{\text{expected feature values induced by current } \omega} \right)$$

contrast to MMP : $f_{\xi_D} - f_{\xi^*_{\omega_i}}$

noisy demonstration version of MMP – ξ not perfect noisy demonstration version of IRL – $\mathcal{D}$ not perfect

---

commonly presented as:

$$\max_{w} \underbrace{\min_{\pi} \left(\lambda H(\pi) - \mathbb{E}_{s,a\sim\pi}\left[\overbrace{w^T f(s,a)}^{\text{cost}} \right] \right)}_{\text{cost of RL on } w \text{ (+entropy)}} \underbrace{- \mathbb{E}_{s,a\sim\pi_D}\left[w^T f(s,a) \right]}_{\text{cost of } \pi_D}$$