1) Max Entropy IRL

Finite Horizon MDP

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

unknown & unobserved

expert dataset: $\mathcal{D} = \{s_i^*, a_i^*\} \sim d_\mu^{\pi^*}$

State/action distributions

$P_h^\pi(s,a;\mu)$ = probability of visiting (s,a) at timestep h following P, π

$$d_\mu^\pi(s,a) = \frac{1}{H}\sum_{h=0}^{H-1} P_h^\pi(s,a;\mu)$$

$$d_\mu^\pi(s) = \sum_{a\in\mathcal{A}} d_\mu^\pi(s,a)$$

Linear Rewards

$$r(s,a) = \theta_*^T \phi(s,a) \qquad \theta_* \in \mathbb{R}^d \quad \phi: \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$$

unknown known

$$\mathbb{E}_{s,a\sim d_\mu^\pi} \phi(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^*}} \phi(s,a) \Rightarrow \mathbb{E}_{s,a\sim d_\mu^\pi} r(s,a) = \mathbb{E}_{s,a\sim d_\mu^{\pi^*}} r(s,a)$$

Max Ent IRL Problem

max ent $\simeq$ min log likihood

$$\min_\pi \mathbb{E}_{s,a\sim d_\mu^\pi}\left[\log(\pi(a|s))\right]$$

estimate from $\mathcal{D}$

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

Using the Lagrange Formulation

$$\min_{\pi} \max_{w \in \mathbb{R}^d} \underbrace{\mathbb{E}_{s,a \sim d^{\pi}_{\mu}}[\log \pi(a|s)] + w^T\Big(\mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}} \phi(s,a)}_{\mathcal{L}(\pi, w)} - \underline{\mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \phi(s,a)}\Big)$$

$$\mathcal{L}(\pi, w) = \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\left[\log \pi(a|s) - w^T\phi(s,a)\right] + w^T \mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}} \phi(s,a)$$

2) Iterative Max-Ent IRL

Initialize $w_0 \in \mathbb{R}^d$

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

$$\pi^t = \arg\max_{\pi} \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}\Big[-\log \pi(a|s) + \overbrace{w_t^T\phi(s,a)}^{f(s,a)}\Big] \leftarrow \text{soft-VI}$$

$$w_{t+1} = w_t + \eta\Big(\mathbb{E}_{s,a\sim d^{\pi^*}_{\mu}} \phi(s,a) - \mathbb{E}_{s,a\sim d^{\pi}_{\mu}} \phi(s,a)\Big) \leftarrow \checkmark$$

Return $\bar{\pi} = \text{uniform}(\pi^0, \ldots, \pi^{T-1})$

$$\searrow \quad \pi(a|s) = \frac{1}{T}\sum_{t=0}^{T-1} \pi^t(a|s)$$

Best Response Step: RL problem with reward $w_t^T\phi(s,a)$ and a policy dependent term $-\log\pi(a|s)$ (exercise)

$$\mathbb{E}_{s\sim\mu}[V^{\pi}(s)] = \mathbb{E}\left[\sum_{t=0}^{H-1} r_t \,\middle|\, r, P, \pi, \mu\right] = H \cdot \mathbb{E}_{s,a\sim d^{\pi}_{\mu}}[r(s,a)]$$

3) Soft Value Iteration

Use Dynamic programming:

$$\arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H-1} r(s_t,a_t) - \log \pi_t(a_t|s_t) \;\middle|\; \begin{array}{l} s_{t+1}\sim P(s_t,a_t)\\ a_t\sim\pi_t(s_t)\\ s_0\sim\mu\end{array}\right]$$

Initialize $V_H^*(s) = 0$

For $h = H-1, \dots, 0$:

1) $Q_h^*(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(s,a)}\left[V_{h+1}(s')\right]$

2) $\pi_h^*(\cdot|s) = \arg\max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_h^*(s,a) - \log\pi(a|s)\right]$

$$= \frac{\exp(Q_h^*(s,a))}{\sum_{a'\in\mathcal{A}} \exp(Q_h^*(s,a'))} \quad \leftarrow \text{"arg" softmax}$$

3) $V_h^*(s) = \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\left[Q_h^*(s,a) - \log\pi_h^*(a|s)\right]$

$$= \log\left(\sum_{a\in\mathcal{A}} \exp\left(Q_h^*(s,a)\right)\right)$$

Contrast π^* V^* with classic RL solution

$$\operatorname{softmax}_a Q_h^*(s,a) \quad \text{vs.} \quad \max_a Q_h^*(s,a)$$

$$\pi(\cdot|s) = \begin{cases} \arg\max_{\rho} \sum_{a \in \mathcal{A}} \rho(a)\left(Q_h^*(s,a) - \log \rho(a)\right) \\ \text{s.t. } \left(\sum_{a \in \mathcal{A}} \rho(a) = 1\right) \end{cases}$$

$$\mathcal{L}(\overset{\in \mathbb{R}^d}{\rho}, \overset{\in \mathbb{R}}{w}) = \sum_{a \in \mathcal{A}} \rho(a)\,\underline{Q_h^*(s,a)} - \underbrace{\rho(a) \log \rho(a)}_{x \log x} + \underline{w\left(\sum_a \rho(a) - 1\right)}$$

$$\frac{\partial}{\partial \rho(a)} \mathcal{L} = Q_h^*(s,a) - \log \rho(a) - \overbrace{\frac{\cancel{\rho(a)}}{\cancel{\rho(a)}}}^{1} + w = 0$$

$$\underline{\rho(a)} = \exp(Q_h^*(s,a))\, \underline{\exp(w-1)} \quad \forall a$$

$$\frac{\partial}{\partial w} \mathcal{L}$$

$$\rho(a) = \frac{\exp(Q_h^*(s,a))}{\sum_{a' \in \mathcal{A}} \exp(Q_h^*(s,a'))}$$