⑧ Perceptual rewards

Perceptual Rewards

IRL

s_t → visual feature activation at time t ⇒ $[s_{1t}, s_{2t}, \cdots]$

τ → $\{s_1, \cdots s_T\}$ // sequence of trajectory.

$$P(\tau) = p(s_1, \cdots, s_T) = \frac{1}{Z} \exp\left(\sum_{t=1}^{T} R_t(s_t)\right)$$; maxEnt model

(R_t: unknown rewards (target))

Boltzmann distribution.

{Dynamic programming ⇐ challenge: How to compute ??}

Now, next state, $s_{t+1} = \begin{cases} f(a_t, s_t) & \text{Deterministic} \\ p(s_{t+1} \mid a_t, s_t) & \text{Probabilistic.} \end{cases}$

Simplifying Assumption,

$$p(\tau) = \prod_{t=1}^{T} \prod_{i=1}^{N} p(s_{it}) = \prod_{t=1}^{T} \prod_{i=1}^{N} \frac{1}{Z_{it}} \exp\left(R_i(s_{it})\right)$$

where ⇒ $R_t(s_t) = \sum_{i=1}^{N} R_i(s_{it})$

Intermediate stage Discovery:

$$p(\tau) = \prod_{t=1}^{T} \prod_{i=1}^{N} \frac{1}{Z_{it}} \exp\left(R_{i g_t}(s_{it})\right)$$

g_t ↓

[index of good step at time t]