DRA vs Max Ent RL

#project #research #RL #max-entropy

MaxEntropy RL

Equations

$$J_{\mathrm{MaxEnt}}(\pi; p, r) = \mathbb{E}_{a_t \sim \pi(a_t|s_t),\, s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \left[\sum_t r(s_t, a_t) + \alpha \mathcal{H}_\pi \left(a_t|s_t\right)\right]$$

$$q_{\mathrm{soft}}(s, a) = r(s, a) + \gamma V_{\mathrm{soft}}(s')$$

$$V_{\mathrm{soft}}(s) = \alpha \log \mathbb{E}_\pi \left[\exp\left(\frac{q_{\mathrm{soft}}(s, \cdot)}{\alpha}\right)\right]$$

$$\pi(a|s) = \frac{1}{Z} \exp\left(\frac{q_{\mathrm{soft}}(s,a) - V_{\mathrm{soft}}(s)}{\alpha}\right)$$

Source of stoachasticity

1. Stochastic policy
    - The policy is a softargmax with temperature α and $q_{\mathrm{soft}}(s, a)$ as arguments
    - The higher the α, the more stochastic the policy gets
2. "Soft" q-values include entropy
    - The soft q-values don't only rely on reward, but they also incorporate the entropy of the policy in that state

Intuition for behavior

Implications for stakes

All else being equal, if the stakes at two states are different, then the one with the higher stakes will have a higher choice probability for the better option since all that matters for the policy entropy is the stakes at the state.

Implications for frequency

All else being equal, a state that is visited more frequently has more of a bearing on the overall reward. If both states had equal policy entropy, marginally reducing the policy entropy in the state would result in a higher gain in reward than the equivalent change in the less frequent state. Hence, at convergence, the state that is visited more frequently would have a higher choice probability for the better option.