Composing Entropic Policies using Divergence Correction

Jonathan J Hunt, Andre Barreto, Timothy P Lillicrap, Nicolas Heess
Compositional Policies

\[
\text{Image CC from https://www.flickr.com/photos/7363531@N05/4179327917} \quad + \quad \text{Image CC from https://www.flickr.com/photos/cyron/37847330/} \quad = 
\]
Problem

Training tasks: \( r_1, r_2 \)

Transfer task: \( r_b = r_1 + r_2 \)
Prior Work

Generalized Policy Improvement (Barreto et al., 2017)

\[
\{\pi_1, \pi_2, \ldots\} \quad Q_b^{GPI}(s, a) = \max_i Q_b^{\pi_i}(s, a)
\]

Compositional Optimism (Haarnoja et al., 2018) with Maximum Entropy RL

\[
Q_b^{CO}(s, a) = Q_{\pi_1}(s, a) + Q_{\pi_2}(s, a)
\]
Maximum Entropy Generalized Policy Improvement

1. **Successor Features**

   \[ \phi = (r_1, r_2) \]
   \[ \psi^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_i \phi_i \right] \]

   \[ Q^\pi_b(s, a) = \psi^\pi(s, a) \cdot (b, 1 - b) \]

2. **Generalized Policy Improvement**

   \[ Q^{GPI}_b(s, a) = \max_i Q^\pi_i(b, s, a) \]

   \[ \pi(a|s) \propto \exp(Q^{GPI}_b(s, a)) \]
Divergence Correction

- Track the discounted, expected Rényi divergence between policies.

\[ Q_b^{DC}(s, a) = Q^{\pi_1}(s, a) + Q^{\pi_2}(s, a) - C(s, a) \]

- This information allows to recover the optimal compositional policy.

\[ Q_b^{DC} = Q_b^* \]
$r_1, r_2 = \frac{3}{4}$

$r_1 = 1 \quad r_2 = 1$

Results