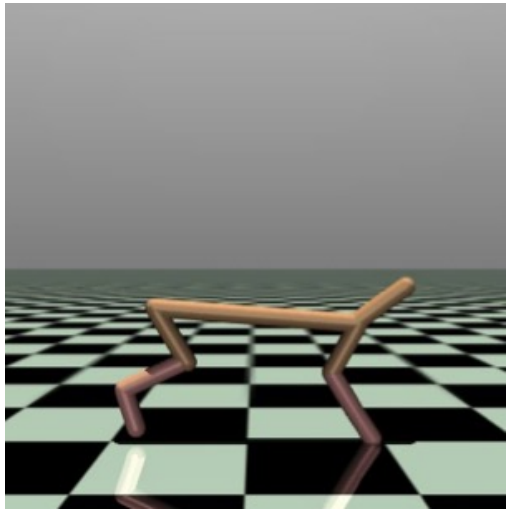
Learning Self-Imitating Diverse Policies [Gangwani et al., 2019]

Challenge:

- Exploration is crucial in parse rewards settings
- Current algorithms don't decompose trajectories to steps

Main ideas:

- $\pi(a|s)$ better reflected by $\rho_\pi\ (s, a)$
- Optimize π by minimizing divergence between ρ_π and ρ_{π^*}
- *Self-imitation from experience*:
  - Approximate ρ_{π^*} by high-rewarded trajectories



Learning Self-Imitating Diverse Policies [Gangwani et al., 2019]

Policy update:

- Approximate gradient of

  Jensen-Shannon divergence