Continuing Control

No episodes! Life goes on forever

Discounting

Continuing Control + function approximation → no discounting in the objective

Not an Optimization Problem

Comparing policies:

\[ v^\pi(s) \geq v^{\pi'}(s) \quad \forall s \]

No representable policy is better for every state!

Policy
Optimal policy (for every state)
Many "optimal" representable policies (each better in some states, worse in others)

Need an Objective Function

How to evaluate policy performance:
- Value of start state?
  ✓ Episodic tasks
  ✗ Continuing tasks (start states do not matter)
- Weighted average over states?
  ✓ Average reward
  ✓ Interest function (but changes problem formulation)

Naively using discounted algorithms should not be the first choice for continuing tasks

✕ Maximizing Discounted Return ✗

Greedily maximizing discounted value does not maximize average reward (Sarsa, Q-Learning)

Optimal policy depends on \( \gamma \)

Small \( \gamma \): go left
Critical \( \gamma^* \approx 0.84 \)
Large \( \gamma \): go right

✕ Increasing \( \gamma \rightarrow 1 \) ✗

Critical \( \gamma \) is unknown and problem-specific

Algorithms become unstable as \( \gamma \rightarrow 1 \)

See paper for details: arxiv.org/abs/1910.02140