Generalizing Curricula for Reinforcement Learning

Sanmit Narvekar and Peter Stone
Department of Computer Science
University of Texas at Austin
{sanmit, pstone} @cs.utexas.edu
Successes of Reinforcement Learning

Approaching or passing human level performance

BUT

Can take *millions* of episodes! People learn this *MUCH* faster
People Learn via Curricula

People are able to learn a lot of complex tasks very efficiently
Curricula in RL

Curricula must be recreated from scratch for each new task or agent

Can we use knowledge gained about learning a curriculum for one task to speed up learning of a curriculum for a new task?
Sequencing as an MDP

Curriculum Task

Task 1
- Environment
- RL Agent
  - Action
  - State Reward
- Curriculum Action

Task 2
- Environment
- RL Agent
  - Action
  - State Reward
- Curriculum Action

Task N
- Environment
- RL Agent
  - Action
  - State Reward
- Curriculum Action

Curriculum Task
- Curriculum Agent
  - Curriculum State
  - Curriculum Reward
Sequencing as an MDP

- **State space** $S^C$: All policies $\pi_i$ an agent can represent
- **Action space** $A^C$: Different tasks $M_j$ an agent can train on
- **Transition function** $p^C(s^C, a^C)$: Learning task $a^C$ transforms an agent’s policy $s^C$
- **Reward function** $r^C(s^C, a^C)$: Cost in time steps to learn task $a^C$ given policy $s^C$
Combining CMDPs with UVFAs

• Universal Value Functions learn a VF over states and goals

\[ v_\pi(s, g) = \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} r_g(s_t, a, s_{t+1}) \bigg| s_0 = s \right] \]

• In our setting, goals are tasks

• For now, we restrict ourselves to navigational tasks, where tasks can be represented by their start and end coordinates

• 2 stream architecture to create an embedding over states and goals, then merge

Schaul et al. (2015)
Experimental Results

• Evaluate whether curriculum policies learned for one set of tasks can generalize to a novel set of unseen tasks

• Navigational tasks
  • Start x
  • Start y
  • End x
  • End y

• 9900 possible tasks

• 8 + 1 source tasks
Interpolation Results

- Randomly **shuffle** all tasks
- Present tasks **one by one**
- Each task seen is novel, though similar tasks might have been seen previously
- Learns to **interpolate** between tasks

![Graph showing cost to learn target task over CMDP episodes with and without curriculum](image-url)
Extrapolation Results

- Split tasks into train/test set

- Test set tasks start in top left room and end in bottom right

- Train on source tasks for 200 episodes, then evaluate on test set

- Learns to extrapolate to unseen types of tasks
Summary

• Curricula often need to be **recreated from scratch** for each new agent or task

• Showed curriculum policies can **generalize** to produce curricula for unseen tasks

• Showed that **tasks can be used as goals** in a UVFA to make this possible

• Extend to non-navigational tasks, where a **more general representation for tasks** is needed