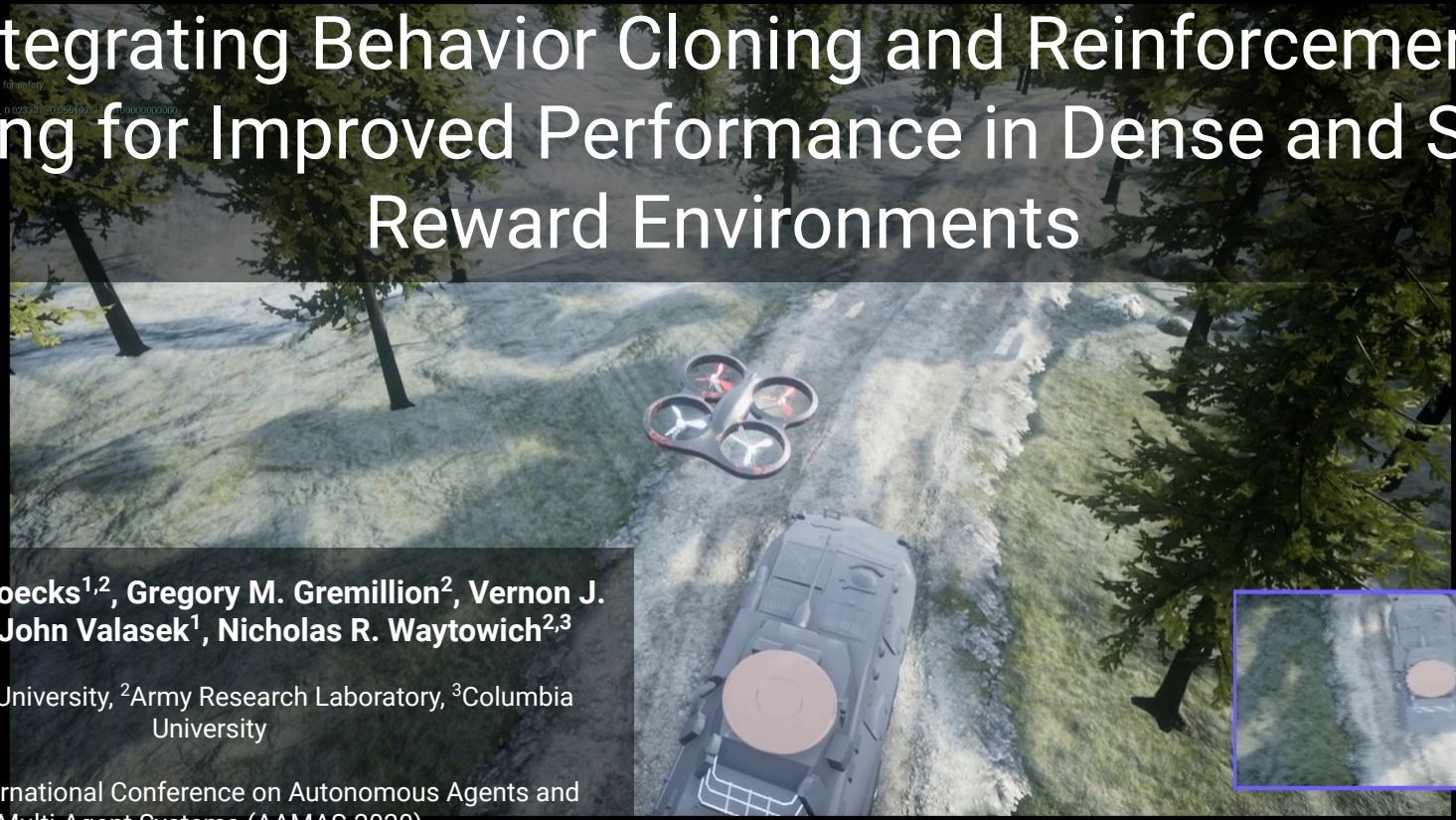
Integrating Behavior Cloning and Reinforcement Learning for Improved Performance in Dense and Sparse Reward Environments



Vinicius G. Goecks[1,2], Gregory M. Gremillion[2], Vernon J. Lawhern[2], John Valasek[1], Nicholas R. Waytowich[2,3]

[1]Texas A&M University, [2]Army Research Laboratory, [3]Columbia University

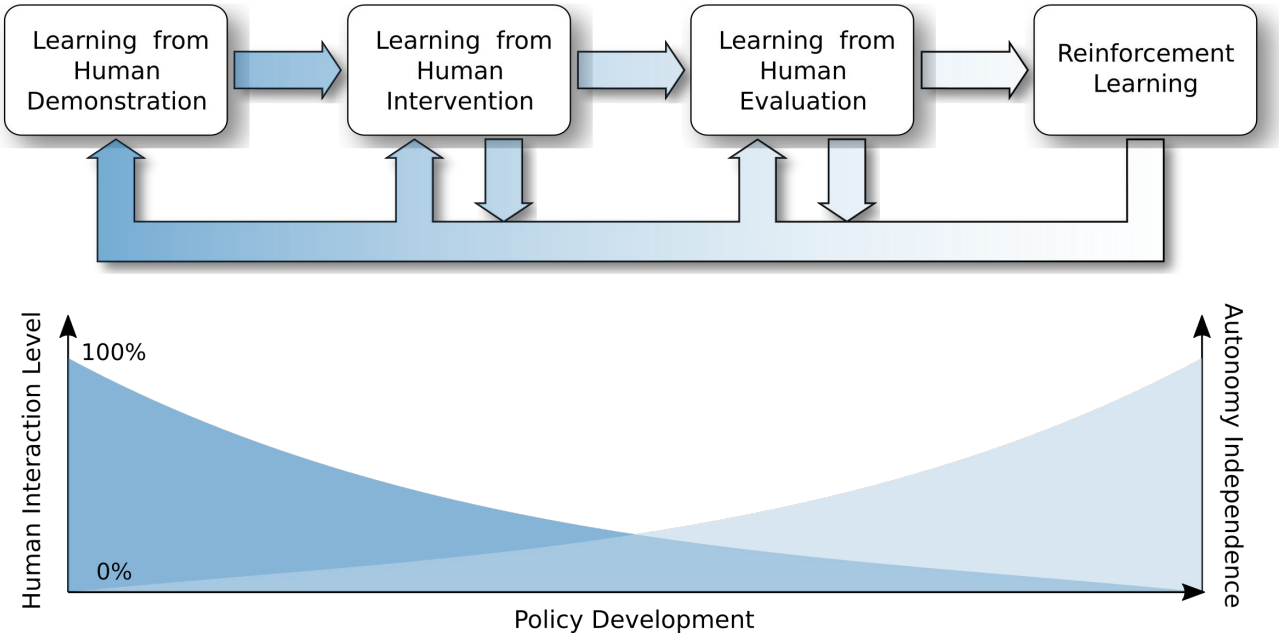
The 2020 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2020)

*Contact information at the end of video.

Vehicle Systems & Control Laboratory
AEROSPACE ENGINEERING

Cycle-of-Learning

A framework for **rapidly training autonomous agents** using a combination of **multiple human interaction modalities** and **reinforcement learning** (self-learning).



1) Nicholas R. Waytowich, Vinicius G. Goecks, Vernon J. Lawhern, "Cycle-of-Learning for Autonomous Systems from Human Interaction", AAAI-FSS AI-HRI, 2018.
2) V.G. Goecks, G.M. Gremillion, V.J. Lawhern, J. Valasek, N.R. Waytowich., "Efficiently Combining Human Demonstrations and Interventions for Safe Training of Autonomous Systems in Real-Time," *AAAI*, 2019.

Key Ideas

1. Pre-training phase
    - Leverages previous task demonstration data to initialize actor (policy) and critic networks.

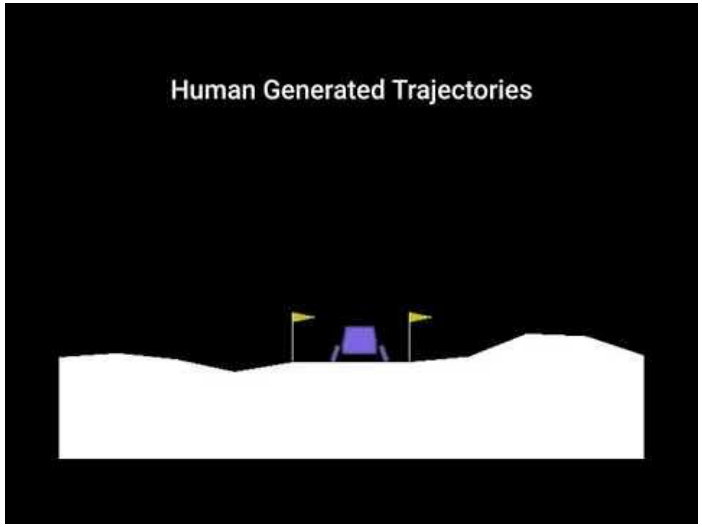
2. Combined loss function
    - Combines supervised and reinforcement learning loss functions in a single update to enable transition between these policies.

3. Component Analysis
    - Investigates the contribution of each loss component and training phase to the overall policy training.

Approach

1. Collect expert demonstrations of the task to be learned:

LunarLanderContinuous-v2
(dense/sparse reward)



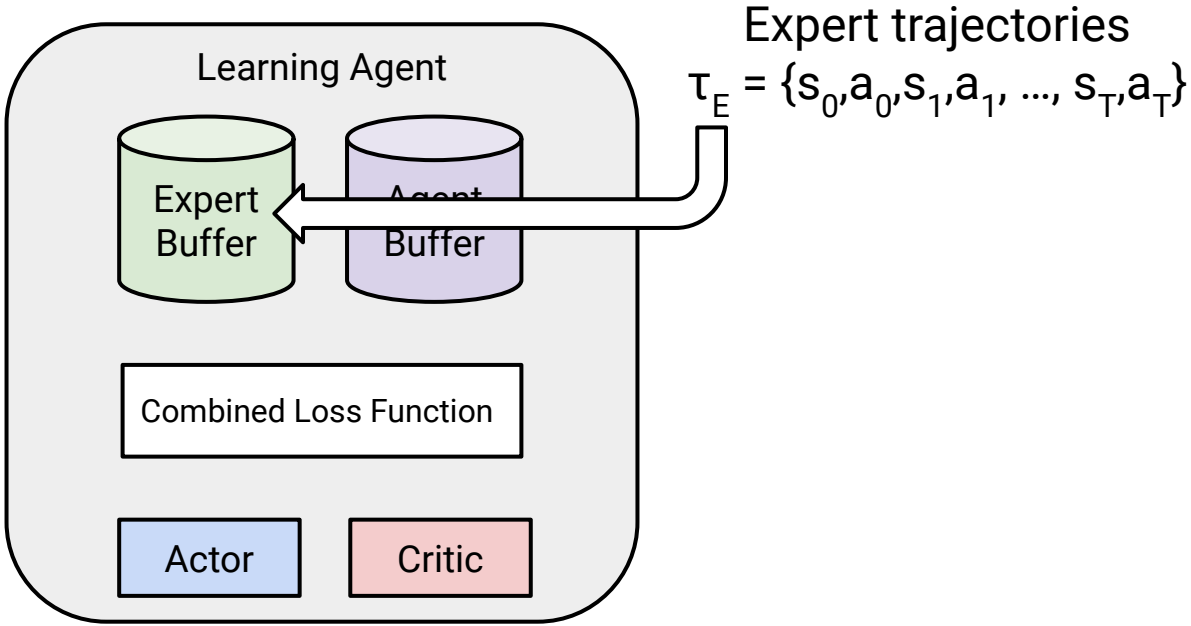
Video: https://youtu.be/8phyJO61nDM

Quadrotor Landing in Microsoft AirSim
(sparse reward)

Video: https://youtu.be/IK_PIHE1be0

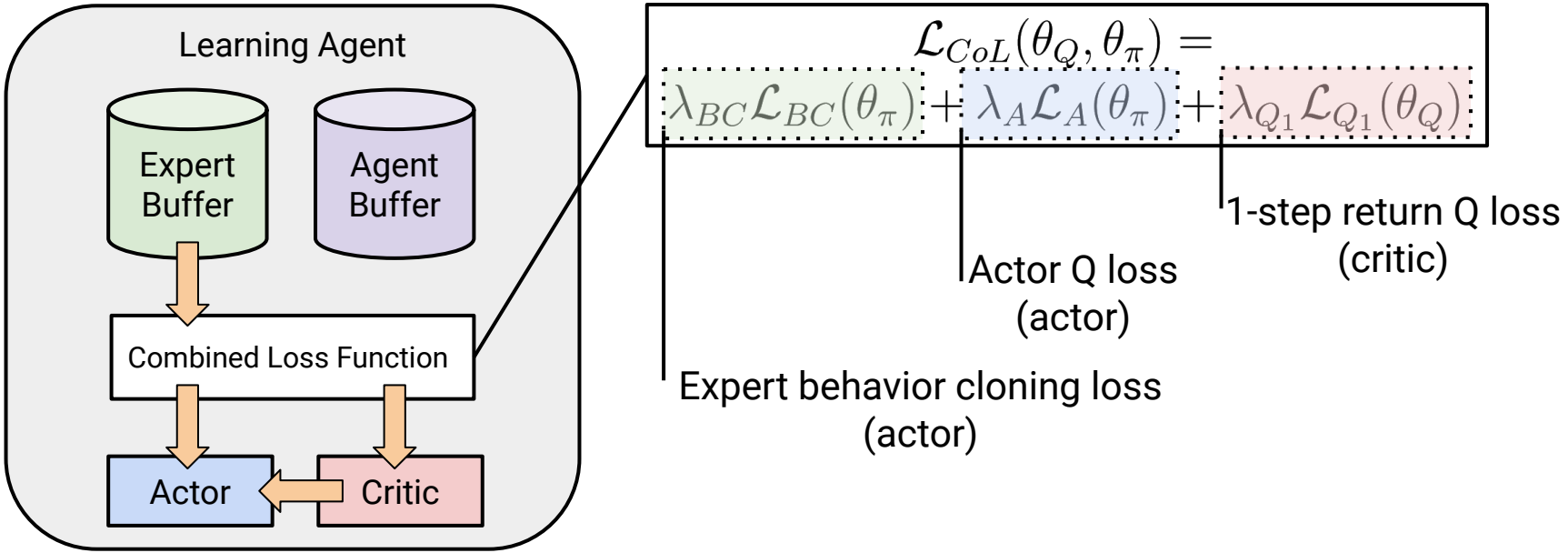
Approach

2. In the learning actor-critic agent, store trajectories in a separate expert buffer:



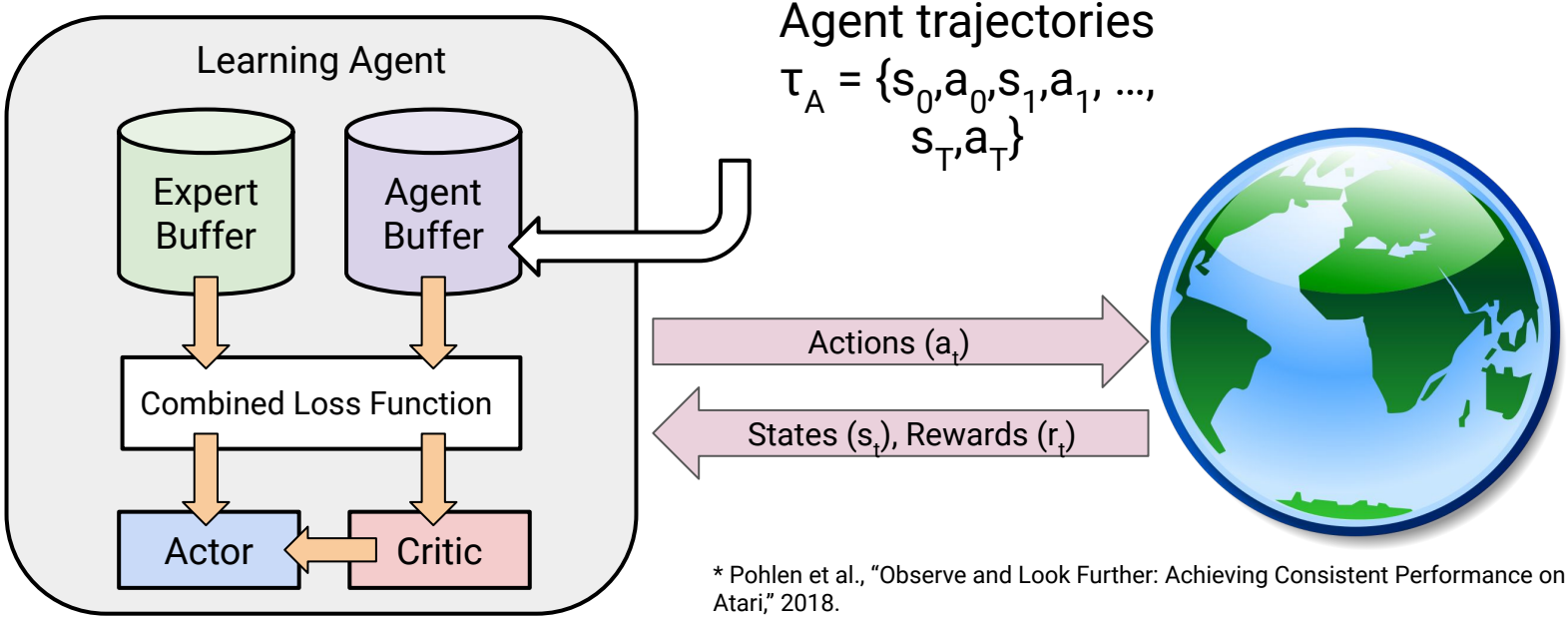
Approach

3. During the pre-training phase, the agent uses the expert data to train both actor and critic networks without interacting with the environment.



Approach

4. During the training phase, the agent interacts with the environment, trajectories are stored in the agent buffer, and actor and critic networks are updated by sampling data from both buffers with a fixed ratio*.



* Pohlen et al., "Observe and Look Further: Achieving Consistent Performance on Atari," 2018.

Qualitative Results

Video: https://youtu.be/0iBY84pI480

Baselines

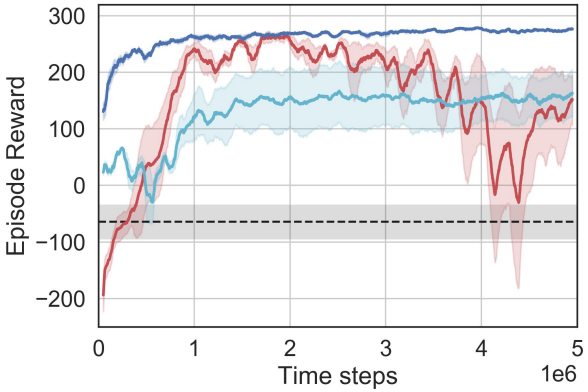
- Behavior Cloning (BC)
- Deep Deterministic Policy Gradient (DDPG)*
- Demo Augmented Policy Gradient (DAPG) **
- Cycle-of-Learning (CoL, *ours*)

* Lillicrap et al., “Continuous control with deep reinforcement learning,” 2015.
** Rajeswaran et al., “Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations,” 2018.

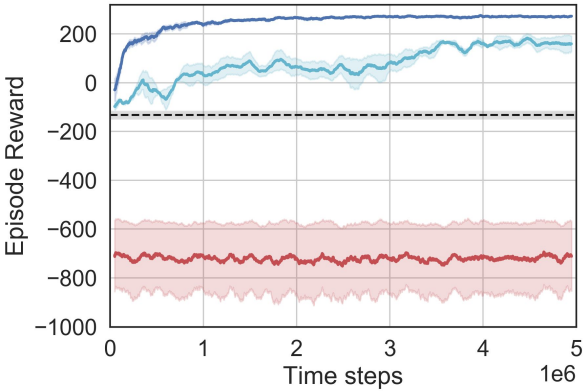
Baselines

- Total episode reward vs training time (for multiple domains, reward functions):
  - Evaluated for **Behavior Cloning (BC)**, **Deep Deterministic Policy Gradient (DDPG)**, **Demo Augmented Policy Gradient (DAPG)**, and **Cycle-of-Learning (CoL, *ours*)**:

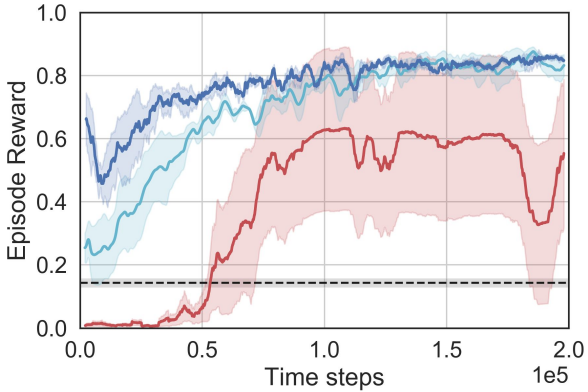
LunarLanderContinuous-v2 (dense reward)



LunarLanderContinuous-v2 (sparse reward)



Landing in Microsoft AirSim (sparse reward)



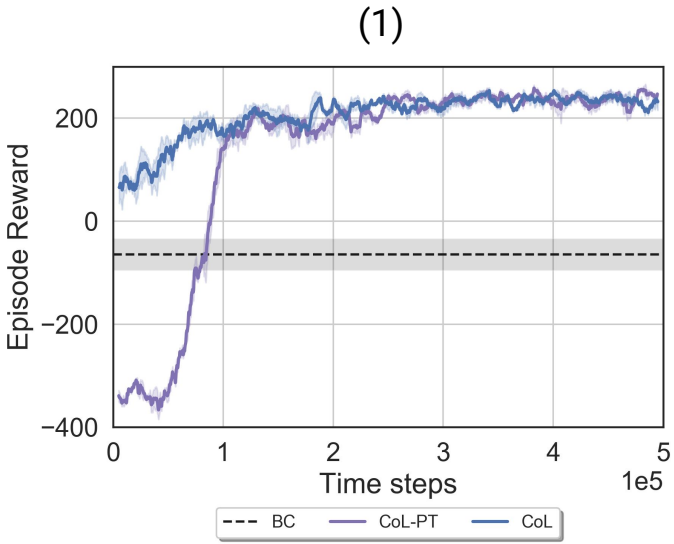
Component Analysis

1. Effects of Pre-Training*:

   a. Assesses the impact on learning performance of not pre-training the agent, while still using the combined loss in the RL phase.

   b. Compares **Behavior Cloning (BC)**, **Cycle-of-Learning without the pre-training phase (CoL-PT)**, and the complete **Cycle-of-Learning (CoL)**.

   c. Highlights that the benefit of pre-training is improved initial response and significant speed gain in reaching steady-state performance level, without qualitatively impacting the long-term training behavior.



* Evaluated on LunarLanderContinuous-v2 environment, dense reward signal.

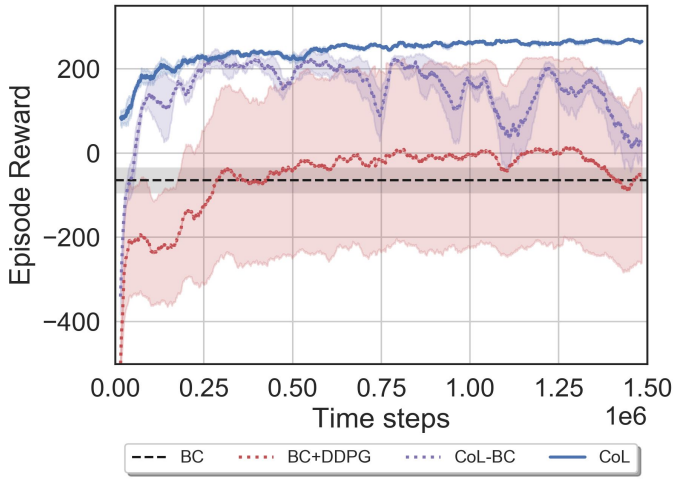
Component Analysis

1. Effects of Combined Loss*:

   a. Assesses the impact on learning performance of the behavior cloning loss component $\mathcal{L}_{BC}$ given otherwise consistent loss functions in both pre-training and RL phases.

   b. Compares **Behavior Cloning (BC)**, **BC followed by DDPG (BC+DDPG)**, **Cycle-of-Learning without the BC loss (CoL-BC)**, and the complete **Cycle-of-Learning (CoL)**.

(2)



   c. Highlights that the BC loss contributes toward maintaining performance throughout training, anchoring the learning to some previously demonstrated behaviors that are sufficiently proficient.

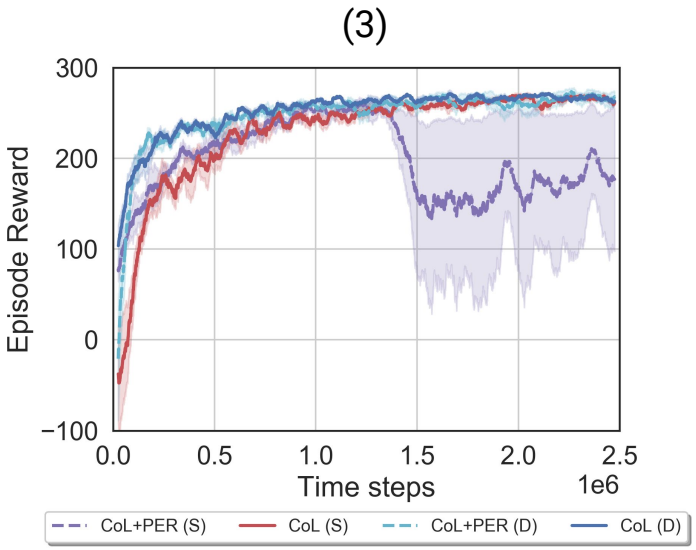
* Evaluated on LunarLanderContinuous-v2 environment, dense reward signal.

Component Analysis

1. Effects of Human Experience Replay Sampling*:

   a. Assesses the impact of mixing agent and expert samples, namely, a fixed sampling or following prioritized experience replay (PER) based on the temporal-difference error.

   b. Compares **Cycle-of-Learning with PER (CoL+PER (S))** to **CoL with fixed buffer ratio (CoL (S))** for *sparse rewards*, and compares **CoL with PER (CoL+PER (D))** to **CoL with fixed buffer ratio (CoL (D))** for *dense rewards*.



   c. Highlights that the fixed sampling ratio is a more robust mechanism of incorporating experience data likely because it grounds performance to demonstrated human behavior throughout training.

* Evaluated on LunarLanderContinuous-v2 environment, dense and sparse reward signal.

Summary

This work enables **transition** from behavior cloning to reinforcement learning **without performance degradation** and **improves** reinforcement learning in terms of **overall performance** and **training time**, for both **dense** and **sparse** reward scenarios.

Thank you!

Integrating Behavior Cloning and Reinforcement Learning for Improved Performance in Dense and Sparse Reward Environments

For more information:
https://vggoecks.com/cycle-of-learning/

Vinicius G. Goecks: vinicius.goecks@tamu.edu
Gregory M. Gremillion: gregory.m.gremillion.civ@mail.mil
Vernon J. Lawhern: vernon.j.lawhern.civ@mail.mil
John Valasek: valasek@tamu.edu
Nicholas R. Waytowich: nicholas.r.waytowich.civ@mail.mil

Research was sponsored by the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-18-2-0134. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The 2020 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2020)

U.S.ARMY DEVCOM ARMY RESEARCH LABORATORY

Vehicle Systems & Control Laboratory
AEROSPACE ENGINEERING