REWARD SHAPING FOR GUIDING REINFORCEMENT LEARNING

STUDENT: BAICEN XIAO, NETWORK SECURITY LAB

Motivation

- Reinforcement learning (RL) learns a policy by interacting with environments in order to maximize certain reward
- Sparse/ delayed reward makes learning hard and even distract RL agents from true goal
- Domain knowledge can be incorporated to make the reward signal denser

Research Goals

- Use potential-based methods to learn stochastic policies
- Integrate other form of domain knowledge with RL

Potential-based Reward Shaping

- Potential-based reward shaping (PBRS) modifies reward in a principled manner that allows for the recovery of optimal policies
- Potential-based reward $F(s_t, s_{t+1}) = \emptyset(s_{t+1}) - \emptyset(s_t)$
- Focus limited to deterministic policies, discrete envs.

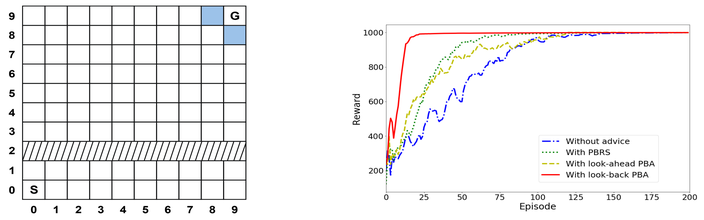
SOLUTION: Generalize potential-based methods to stochastic policies and continuous state/ action spaces

- **AC-PBA** : algorithm on actor-critic architecture augmented with potential-based advice

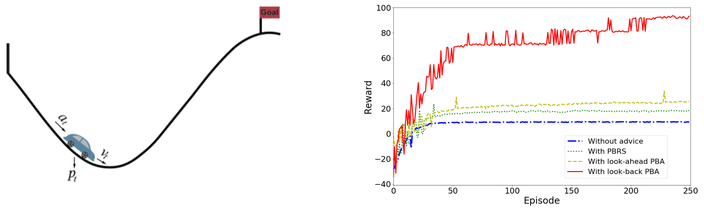
Contributions on PBRS

- An algorithm, AC-PBA, describing an advantage actor-critic architecture augmented with potential-based advice (PBA)
- Guarantees on AC-PBA's convergence.

Validate AC-PBA in Two Environments



Puddle-jump: convergence to target 5 times faster and higher reward with AC-PBA for more difficult tasks



Mountain-car: optimal solution in 100% of trials with AC-PBA

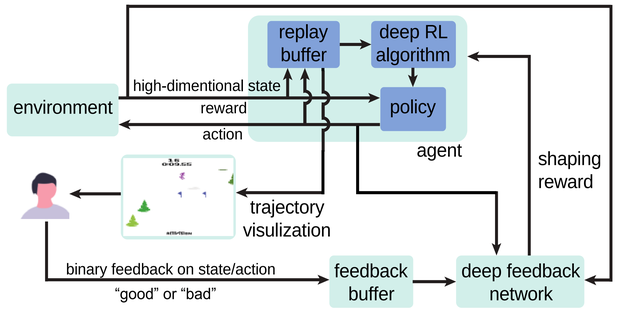
Interactive Reward Shaping

- Potential-based reward shaping (PBRS) may not be always available, especially in high dimensional state/ action spaces
- A human player is easily able to play and win games in setups where the reward structure is sparse or significantly delayed, while deep RL algorithms struggle.

SOLUTION: Feedback-based Reward Shaping (FRESH) in high-dimensional spaces with sparse/ delayed rewards

- **MECHANISM:** feedback neural network to effectively represent human feedback & predict model uncertainty
- At each state in the trajectory, the operator indicates whether the action taken in that state is good or bad
- A deep feedback neural network is used to allow the deep RL algorithm to generalize feedback signals (also be able to predict model uncertainty) obtained during training to unseen states and actions at test-time

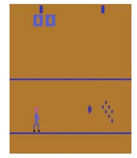
FRESH Framework

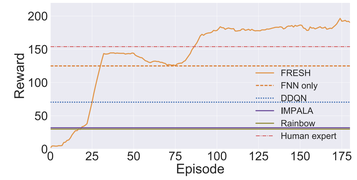


Schematic for FRESH (Feedback-based Reward SHaping)

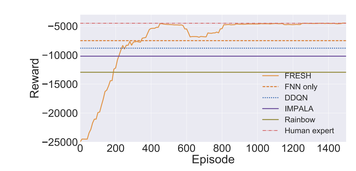
Validate FRESH in Two Atari Games

- FRESH outperformed deep RL algorithms in *Bowling* (**187 vs. 70.5**) and *Skiing* (**-4400 vs. -9000**) Atari games





- FRESH outperformed expert human in *Bowling* by **21.4%**



References

[1] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in International Conference on Machine Learning (ICML), 1999..
[2] B. Xiao, B. Ramasubramanian, A. Clark, H. Hajishirzi, L. Bushnell, and R. Poovendran, `Potential-based Advice for Stochastic Policy Learning', in Proc. Conference on Decision and Control 2019
[3] B. Xiao, Q. Lu, B. Ramasubramanian, A. Clark, L. Bushnell, R. Poovendran, "FRESH: Interactive Reward Shaping in High-dimensional State Spaces Using Human Feedback", International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2020.

Sponsors

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISORS: RADHA POOVENDRAN, LINDA BUSHNELL

SPONSORS: OFFICE OF NAVAL RESEARCH, ARMY RESEARCH OFFICE