Exploring Value Function Transfer Between On-Policy and Off-Policy Methods in Tabular Gridworld

Can Kocak , Paul Steinbrink
Leibniz University - Institute of Artificial Intelligence

Scan me for GitHub-Repository

1 TL;DR

- Value function transfer between SARSA (on-policy) and Q-Learning (off-policy).
- Using two different exploration strategies ε-greedy and softmax.
- Cross-paradigm transfer often leads to trade off between return reward and training error.

2 Motivation & Problem Setting

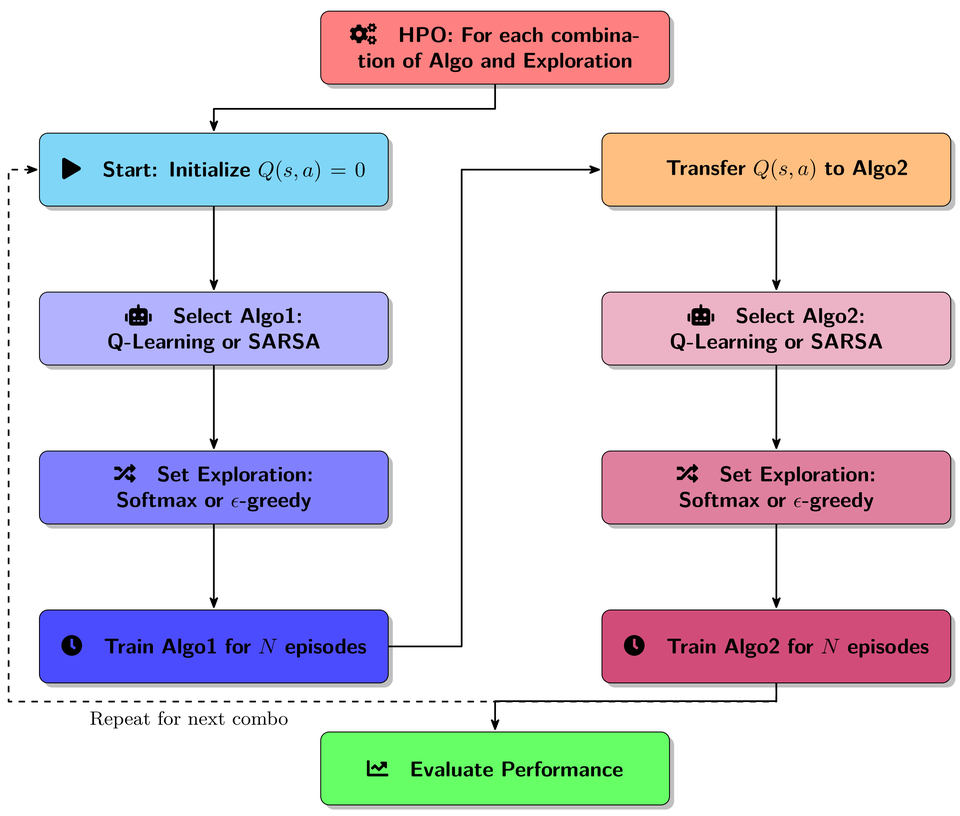
Motivation

- Transferring value function can accelerate learning in tabular RL.
- Exploration strategies like ε-greedy and softmax may influence transfer success.

Problem Setting

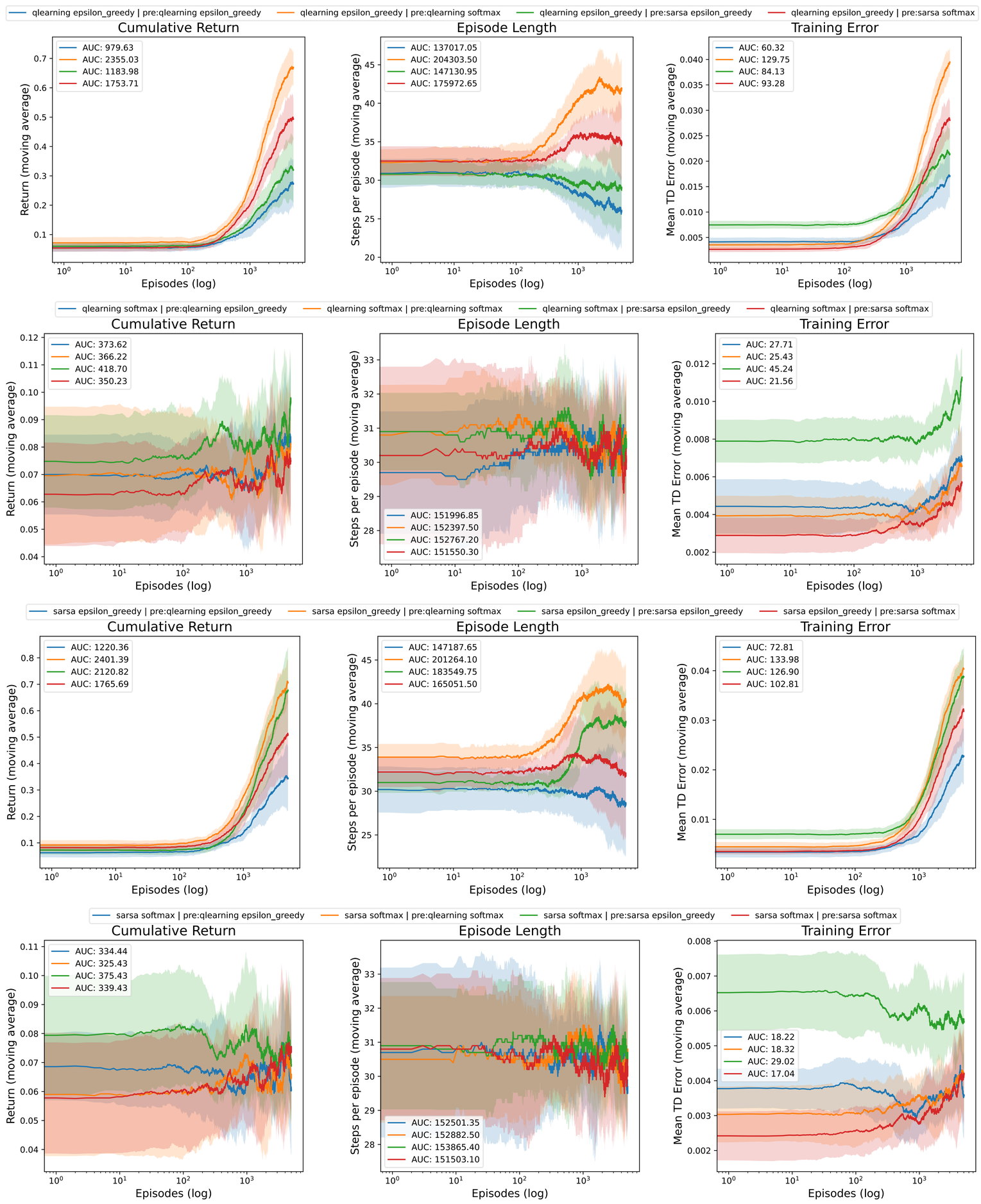
- Can value functions be effectively transferred between off- and on-policy tabular RL methods?
- How does the exploration strategy (ε-greedy vs. softmax) influence such transfer?

3 Approach



- FrozenLake Environment with a grid-size of 7x7.
- Transferring Q-Table.
- 5000 Episodes per run for pre- and transfer-training.
- Resetting Policy after transfer (e.g. ε-greedy).

4 Key Insights



- Off- to On-Policy: Sharp Q-values transferred, but overestimation needs correction.
- On- to Off-Policy: Conservative values get overwritten fast.
- ε-greedy to softmax: Only if Q-values ranked well. Otherwise, it causes noisy probabilities.
- softmax to ε-greedy: Smooth values confuse greedy selection.

5 Future Works

- Explore more complex environments.
- Test additional RL algorithms and exploration strategies.
- Exploring the optimal timing or scope of transfer.
- Explore generalization with switching the environment.