Solving the Harvest CPR Appropriation Problem with Policy Gradient Techniques

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

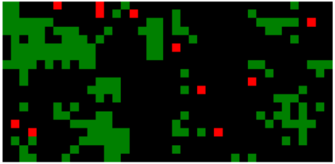
AAS Final Project - Academic Year 2020/2021

Alessio Falai
alessio.falai@studio.unibo.it

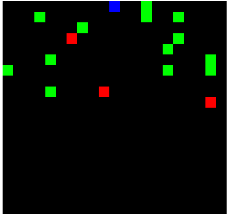
September 9, 2021

Alma Mater Studiorum - University of Bologna

Environment



Full environment



Local observation

- Small 25 × 7 grid for the single-agent setting and big 39 × 17 map for multi-agent scenarios
- 9 actions in total: movement + tagging + gifting
- Local observation: RGB image of size 3 × 20 × 21 (20 squares ahead and 10 squares on each side of the agent)

Social Learning

- Social Learning: reshape the reward function of other agents with the goal of promoting cooperation
- Gifting: peer-rewarding strategy in which agents can reward others with a new specialized action
- Gifting mechanisms: each time an agent sends a gift g, its gifting budget is decremented by g
  - Zero-Sum: the budget is infinite, but the agent incurs a penalty $-g$ for every gifting action taken
  - Fixed Budget: the budget is fixed at the start of the episode and when it's empty no more gifting can happen
  - Replenishable Budget: the budget expands as a function of collected environmental rewards

Experiments

1. Single-agent DQN vs VPG with RLlib
2. Custom VPG vs TRPO vs PPO on Cartpole
3. Custom VPG vs TRPO vs PPO on single-agent Harvest
4. Custom PPO on multi-agent Harvest, with and without Zero-Sum gifting
5. Custom PPO on multi-agent Harvest, with Replenishable and Fixed Budget gifting

Thank you for your attention

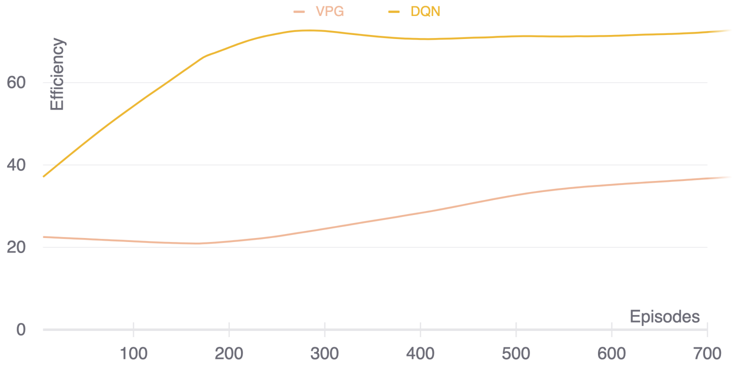
References

- A. Lupu and Doina Precup.
  Gifting in multi-agent reinforcement learning.
  In *AAMAS*, 2020.

- Julien Pérolat, Joel Z. Leibo, Vinícius Flores Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel.
  A multi-agent reinforcement learning model of common-pool resource appropriation.
  CoRR, abs/1707.06600, 2017.

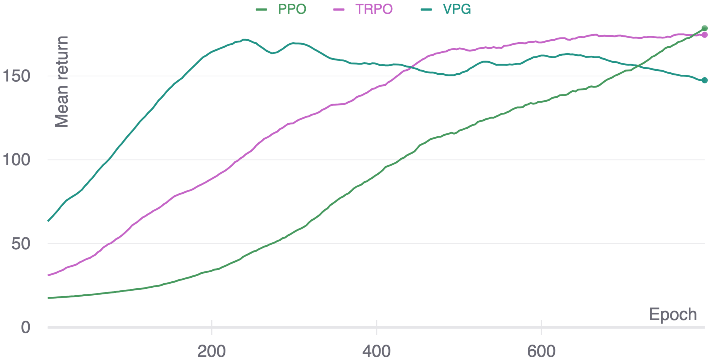
Backup frames

Single-agent DQN vs VPG with RLlib



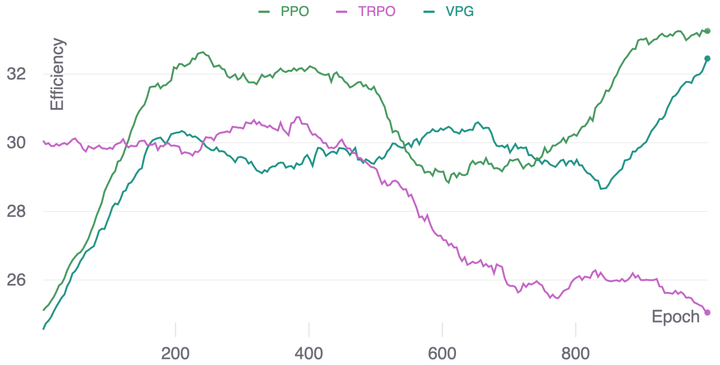
Value-based methods seem more suited for the Harvest environment (higher returns)

Custom VPG vs TRPO vs PPO on Cartpole



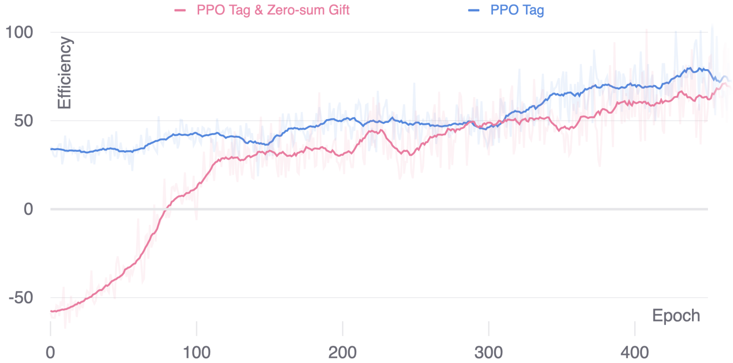
Custom implementations of policy gradient methods are valid, as all agents converge to good returns in the selected test environment

Custom VPG vs TRPO vs PPO on single-agent Harvest



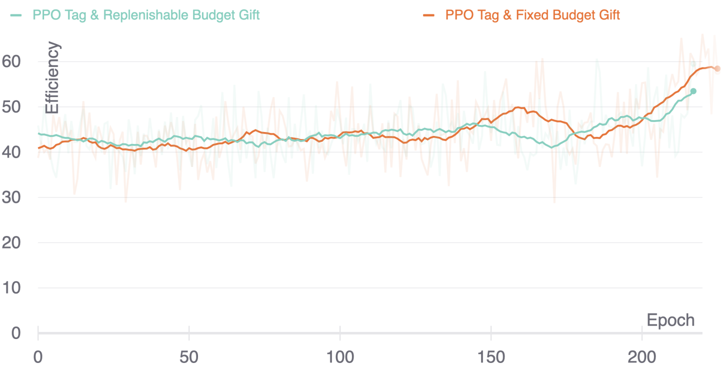
VPG and PPO converge to similar results, while TRPO diverges on the single-agent setting of Harvest

Custom PPO on multi-agent Harvest, with and without Zero-Sum gifting



Agents tend to be very generous at the beginning, while later training stages show that enabling or disabling Zero-Sum leads to similar results

Custom PPO on multi-agent Harvest, with Replenishable and Fixed Budget gifting



Results show that the Replenishable and Fixed Budget gifting strategies tend to follow similar training curves and converge to comparable results