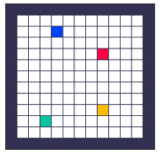
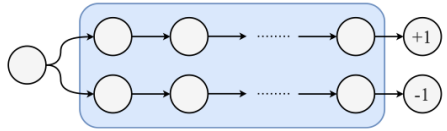
# Discovery of new RL algorithms [OHC+21]

# Key Aspects and Methodology [OHC+21]

[OHC+21]



(a) Grid world

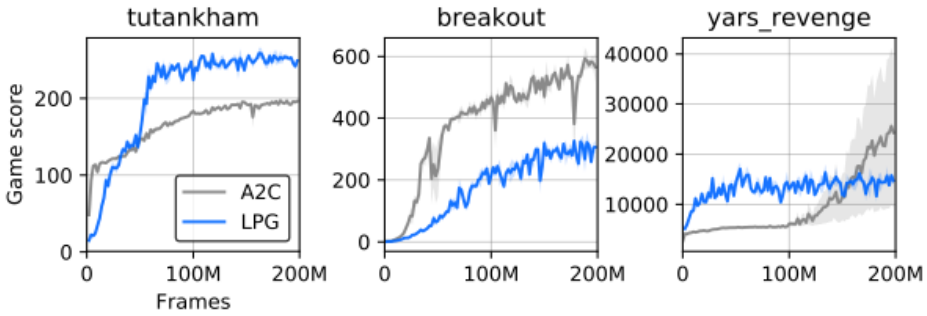


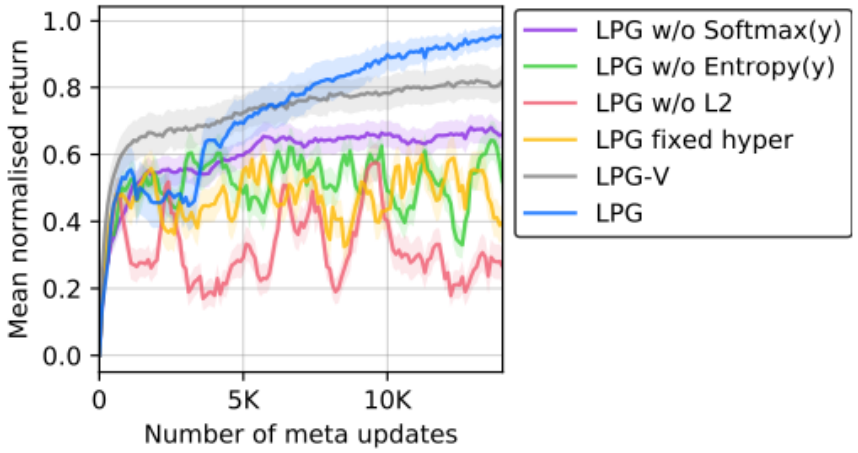

(b) Delayed chain MDP

- **Key Concepts:** Automation of update rule discovery, creating more efficient algorithms which are capable of adaptation to new environments
- **Learned Policy Gradient (LPG):** meta-learning framework that discovers entire update rules
- **Generalization:** LPG trained on toy environments can generalize to more complex tasks and demonstrates robustness and adaptability of this approach
- **Methodology:** LPG trained on various simple environments (e.g. grid worlds, delayed chain MDP)
  - Reward by collecting objects or determined by first action taken
- **Architecture:** LPG is a backward LSTM, produces updates for the agents policy and prediction vectors
- **Meta-Learning Process:**
  - **Inner Loop:** Agents learns within environment
  - **Outer Loop:** Meta-learner adjusts the update rules based on performance across learned environments

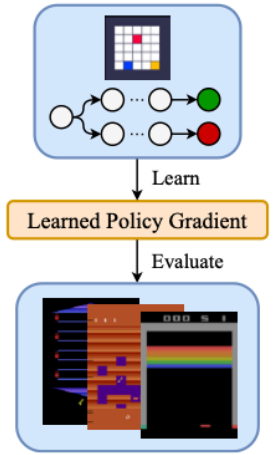
# Experimental Results [OHC+21]

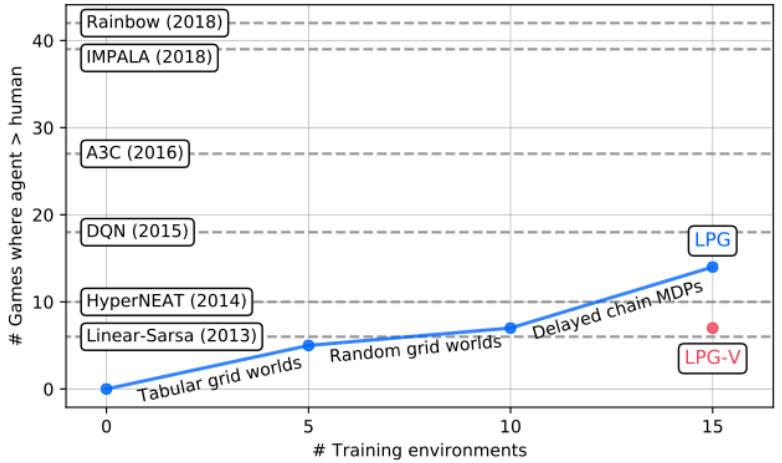
- **Ablation Study:** LPG outperforms all variations and baseline
- **Example Atari learning curves:**




- **Comparison so SOTA algorithms:**



# Sources

[OHC+21] https://arxiv.org/pdf/2007.08794