## The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games

## Paper

* <https://arxiv.org/pdf/2103.01955.pdf>

## Introduction

* Other multiagent algorithms: MADDPG, value decomposed q learning
* Revisits Proximal Policy Optimization (PPO) for multi agent settings

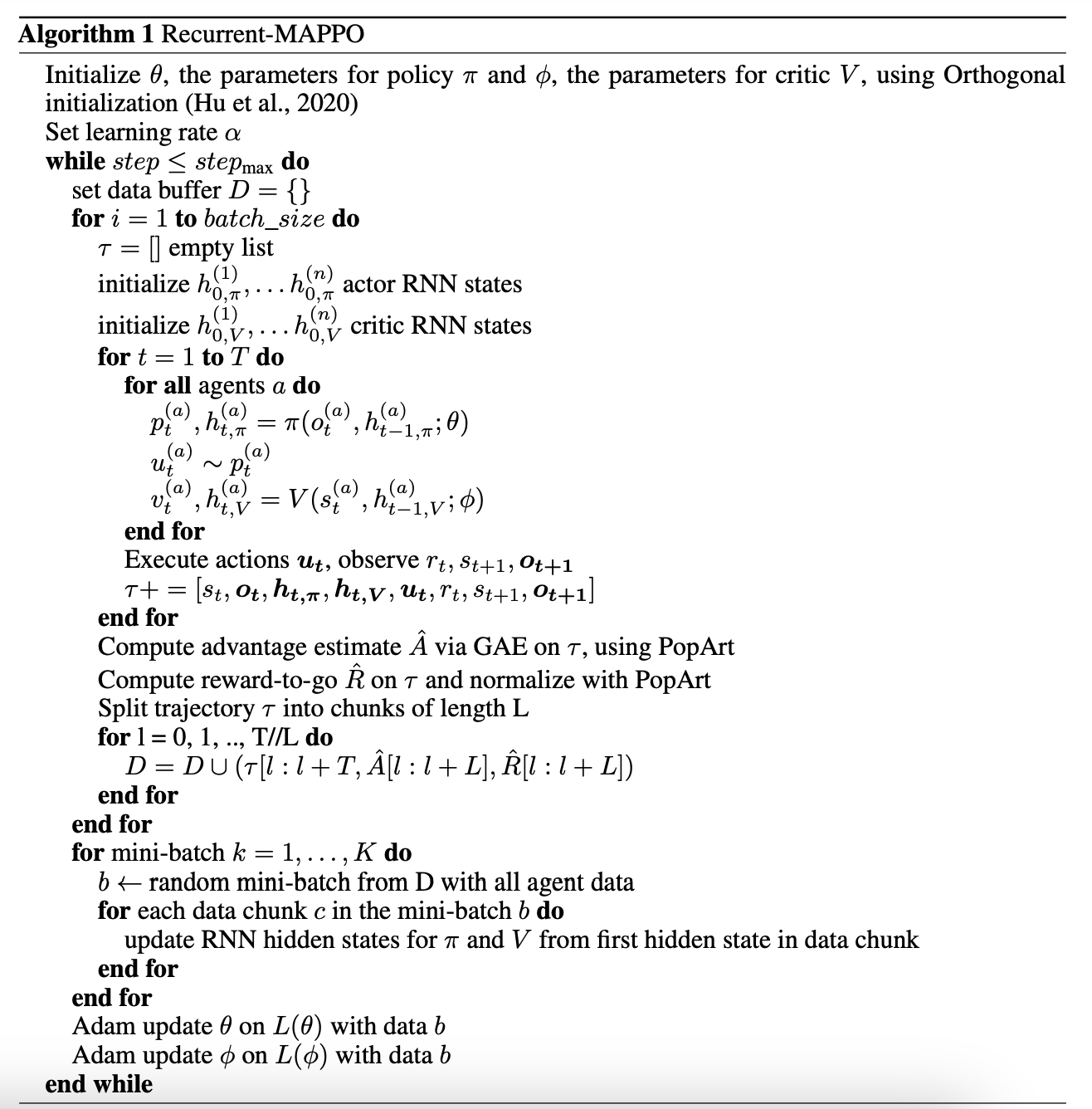
## Related work

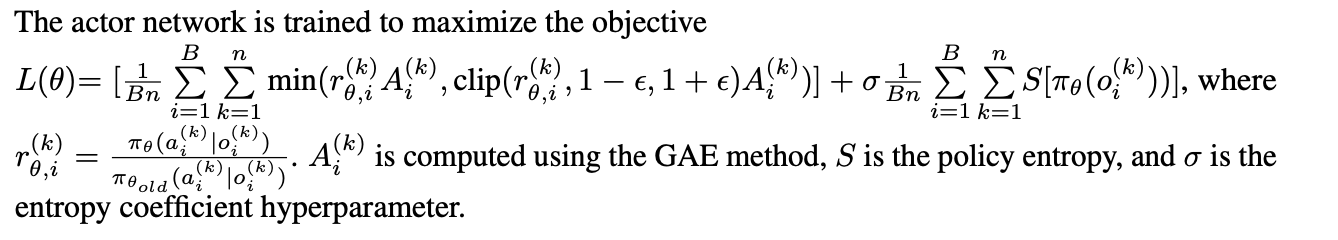
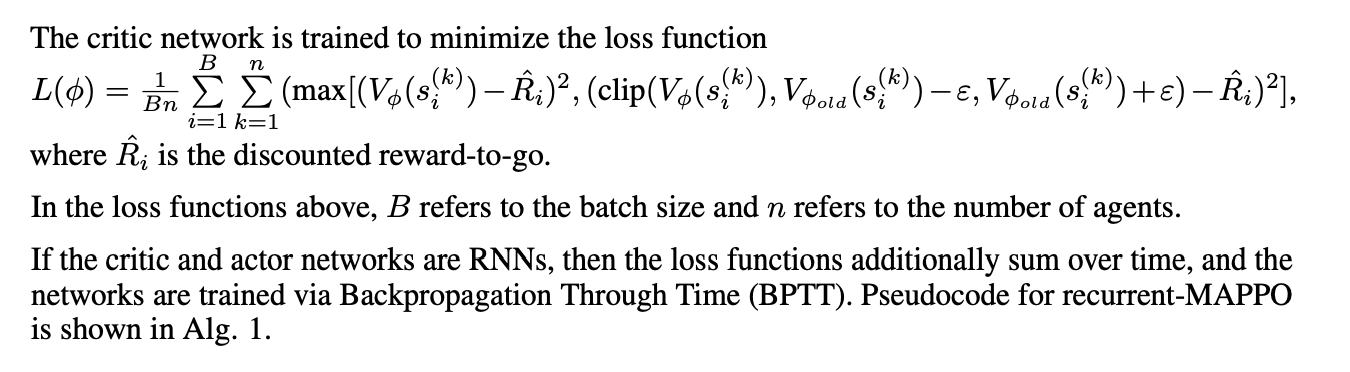
* MARL algorithms
  + Centralized learning - directly learn single policy to produce joint actions of all agents
  + Decentralized learning - each agent optimizes its reward independently
  + Centralized Training and Decentralized execution - in between
    - Value decomposition methods

## Preliminaries

* Decentralized partially observable Markov decision processes (DEC-POMDP) with shared rewards
* <S, A, O<, R, P, n, 𝛾>
  + S: state stpace
  + A: shared action space for each agent i
  + O\_i = O(s;i): local observation for each agent i at global state s
  + P(s’|s , A)
  + A= (a1… an\_
  + R(s,A)shared reward function
  + 𝛾: discount factor r
  + Agents use policy

## Mappo Details



* Mappo trains 2 separate neural networks
  + Actor network with parameters
  + Value function network (critic) with parametsrs
* 
* 
* Trained via backpropogation through time <https://d2l.ai/chapter_recurrent-neural-networks/bptt.html>
* Generalized Advantage Estimation

## Facors influential to ppo performance

* Value normalization
* Input representation to value function
  + “the fundamental difference between many multi-agent CTDE PG algorithms and fully decentralized PG methods is the input to the value network”
  + Options
    - Concatenation Of local observations
    - Environment provided global state
    - Agent specific global state
    - Feature pruned agent specific global state
* PPo clipping
  + PPO clipped importance ratio and value loss
  + Larger clipping values means larger updates to policy and value function ‘
  + Small clipping values → agent’s policy less likey to change per episdoe
* Ppo batch size
  + Shouldnt trian it for too long

## Parameter sharing

* All agents share the same networks
* Recurrent data chunk length

## Death masking

* When an agent dies these agent specific features become zero

<https://arxiv.org/abs/2304.09870>

Main takeaways