Toward Finding Strong Pareto Optimal Policies in Multi-Agent Reinforcement Learning

Le Bang Giang and **Ta Viet Cuong**
University of Engineering and Technology, VNU
HMI Laboratory

Background

- Recent development of MARL mainly focuses on zero-sum games and fully cooperative environments. While the problem of finding *Pareto optimal* policies in *cooperative* MARL remains under-explored.
- A solution is ***strong*** Pareto optimal if no agent can increase their rewards without diminishing those of the others.
- A solution is ***weak*** Pareto optimal if all agents cannot simultaneously increase their rewards.

Motivations

- In fully cooperative environments, globally optimal policies can be found if agents can *communicate*.
- Without communication, only Nash Equilibria are guaranteed.

  👉 *Can current MARL methods find strong Pareto optimal policies in cooperative environments, given arbitrary communication?*

  No,

- In Fig. 1, agent 1 and 2 cannot earn their rewards through their *own* actions.
- Both Policy Gradient and value-based methods can fail to find optimal solutions.

		Player 2 A	Player 2 B
Player 1	A	$1, 2$	$0, 2$
Player 1	B	$1, 0$	$0, 0$

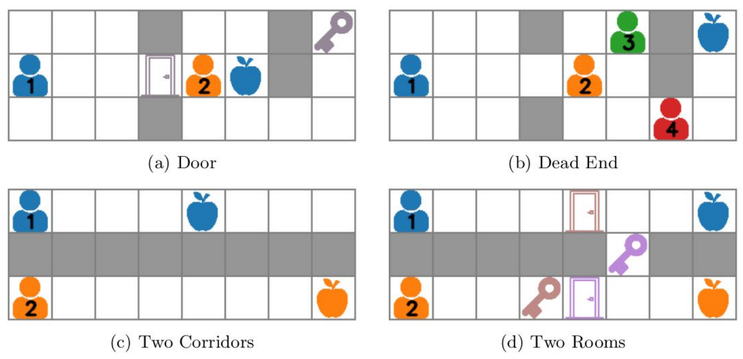
Fig. 1: Example of the matrix game, the tuple in each table cell contains the reward of agent 1 and 2, respectively.

- The problem is that each agent *selfishly* optimizes only their own reward, so no learning can happen if their actions and rewards are uncorrelated.
- Can be fixed straightforwardly: transition from self-centered reward optimization to considering the objectives of other agents

  👉 *Multi-Objective Optimization*

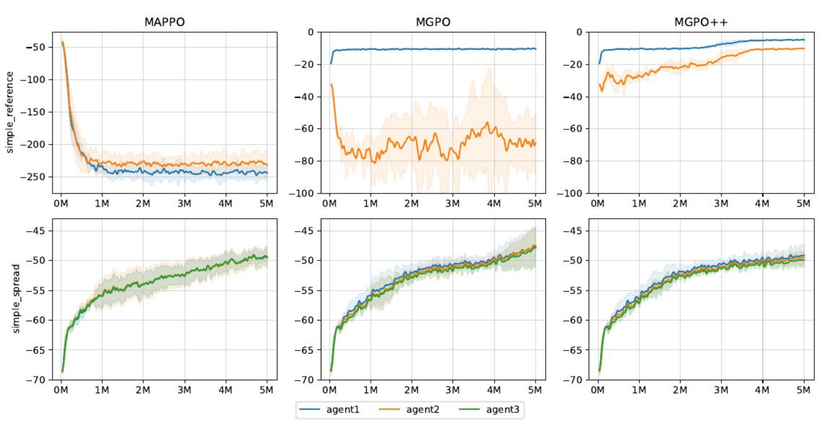
Experiments

- We conduct experiments on two benchmarks: *Gridworld* and *MPE*.
- We compare with several baselines, including independent learning setups: IPPO and IQL, and Centralized training method MAPPO. MGDA++ with MAPPO are denoted as **MGPO++.**



Scenario	agent	MGPO++	MGPO	MAPPO	IQL	IPPO
Door	1	**10.0 ± 0.0**	9.9 ± 0.0	0.0 ± 0.0	−0.0 ± 0.0	−0.0 ± 0.0
	2	**−0.0 ± 0.0**	−6.9 ± 4.9	**−0.0 ± 0.0**	−0.1 ± 0.0	**−0.0 ± 0.0**
Dead End	1	**3.3 ± 4.7**	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.4	0.0 ± 0.0
	2	**−0.0 ± 0.0**	−14.4 ± 0.8	0.0 ± 0.0	−0.0 ± 0.0	0.0 ± 0.0
	3	**−0.0 ± 0.0**	−13.9 ± 2.2	**−0.0 ± 0.0**	−0.1 ± 0.0	**−0.0 ± 0.0**
	4	−0.0 ± 0.0	−14.7 ± 2.2	**0.0 ± 0.0**	−0.1 ± 0.0	−0.0 ± 0.0
Two Corridors	1	**10.0 ± 0.0**	**10.0 ± 0.0**	**10.0 ± 0.0**	9.9 ± 0.0	9.9 ± 0.0
	2	**10.0 ± 0.0**	−1.1 ± 1.6	**10.0 ± 0.0**	9.9 ± 0.0	9.9 ± 0.1
Two Rooms	1	**10.0 ± 0.0**	**10.0 ± 0.0**	−0.0 ± 0.0	−5.6 ± 8.0	6.5 ± 3.6
	2	**9.8 ± 0.2**	−17.7 ± 0.9	0.0 ± 0.0	−0.1 ± 0.0	−0.0 ± 0.0

- MGPO++ can find high rewards for all agents, while MGPO can only find good policy for one agent and get suck at weak Pareto solutions. Other methods fail when cooperations are required.
- Similar results are observed in MPE. MGPO stops learning after one agent converge.
- When rewards are almost fully cooperative, all PG methods converge at the same rate.



Method

- We combine **MGDA** with a Trust region Policy Optimization **MAPPO** (denote **MGPO**). The objective for optimizing the policy *i* with respect to the reward *j* with PPO is

$$L_{\text{PPO}}^{i,j} = \min\left(\frac{\pi_{\text{new}}^{i}}{\pi_{\text{old}}^{i}}\hat{A}_{j}^{\pi_{\text{old}}^{i}}, \text{clip}\left(\frac{\pi_{\text{new}}^{i}}{\pi_{\text{old}}^{i}}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{j}^{\pi_{\text{old}}^{i}}\right),$$

- However, MGDA is known to converge to weak Pareto optimal solutions.
- We give a sufficient condition for a solution to be Strong Pareto optimal in *convex* settings.

Lemma 3. *Under assumption 1, if there exists a convex combination of the subset of all non-zero gradient vectors*

$$\sum_{i\in S}\lambda_i\nabla F_i(x) = 0; \quad \lambda_i > 0, \quad \|\nabla F_i(x)\| > 0 \quad \forall i \in S \tag{4}$$

with $S \subseteq [n], S \neq \emptyset$, then x is Pareto optimal.

Consider non-zero gradient !

- We propose **MGDA++**; Key idea: remove small-norm gradients.

Algorithm 1 MGDA++ Algorithm

Input: $\epsilon > 0$, initial solution x_0

for $k = 0, 1, \ldots$ **do**
  $S_k \leftarrow \emptyset$
  for $i = 1, \ldots, n$ **do**
    Calculate $\nabla F_i(x_k)$
    if $\|\nabla F_i(x_k)\| > \epsilon$ **then** ← filter small gradients
      $S_k = S_k \cup \{i\}$
    end if
  end for
  if $S_k = \emptyset$ **then**
    Stop
  end if
  Find $\{\lambda_i\}_{i\in S_k}$ by solving (3) on the subset of gradients $\{\nabla F_i(x_k)\}_{i\in S_k}$
  $d_k \leftarrow \sum_{i\in S_k}\lambda_i\nabla F_i(x_k)$
  Choose step size t_k
  $x_{k+1} \leftarrow x_k - t_k d_k$
end for

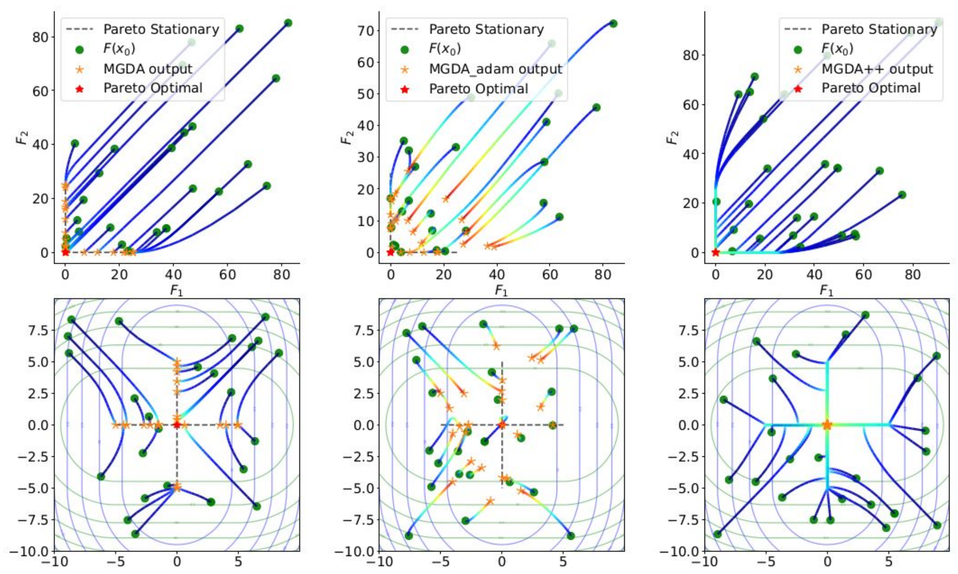
- We give convergence result with two tasks.

Theorem 5. *Under assumptions 1 and 2, for any $\varepsilon > 0$, with $n = 2$ and by choosing $\epsilon < \sqrt{2L\varepsilon}$ and with appropriate choices of each update steps t_k as*

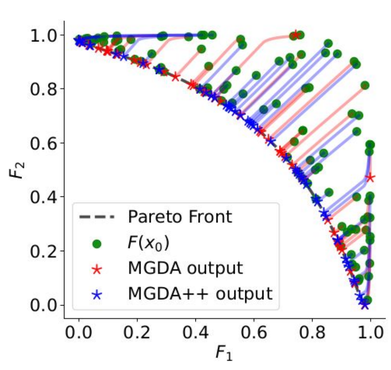
$$t_k = \begin{cases} \max\left(\dfrac{|S_k|\|d_k\|^2 + \left\langle\sum_{i\in\overline{S}_k}\nabla F_i(x_k), d_k\right\rangle}{nL\|d_k\|^2}, 0\right) & \text{if } \|d_k\| > 0 \\ 0 & \text{if } \|d_k\| = 0 \end{cases},$$

with $\overline{S}_k$ the complement of S_k, then each convergent subsequence of MGDA++ converges to either Pareto optimal or ε-Pareto optimal solutions.

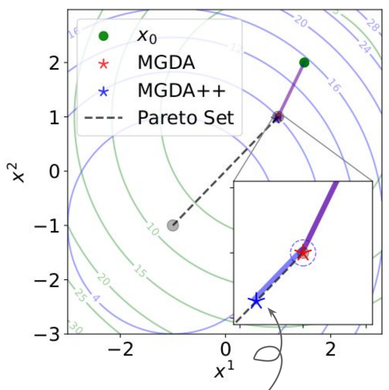
- MGDA and MGDA++ comparisons:



(a) MGDA (b) MGDA w/ Adam (c) MGDA++



both MGDA and MGDA++ behave similarly in normal cases.



MGDA++ rejects small gradient solutions

Address: E3 building, 144 Xuan Thuy road, Cau Giay district, Hanoi, Vietnam
Tel.: +84-24-37547 461. Website: http://www.vnu.edu.vn