Adaptive Learning Rates for Multi-Agent Reinforcement Learning

Professor: Li-Der Chou

Presenter: Chih-Jou Tai

Conference on Autonomous Agents and Multiagent Systems
(AAMAS 2023), May 29 - June 2, 2023, London, UK

Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Introduction

- MARL(Multi-Agent Reinforcement Learning)
- Adaptive Learning Rates: Introduction of adaptive learning rates to cooperative multi-agent reinforcement learning (MARL).
- AdaMa Method:
  - Dynamic Balancing Learning Rates
  - Second-Order Approximation
  - Performance Improvement & Cost Reduction

Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Related Work

	AdaGrad	AdaDelta	RMSprop	WoLF	AdaMa
Learning Rate Adjustment	Accumulated gradients	Moving average of gradients	With decay factor introduced	Adjusts based on equilibrium strategy in games	Dynamically adjusts towards maximally improving Q-values
Training Process	Manual preset learning rate	Less manual presetting required	Less manual presetting required	Requires solving for equilibrium strategy	Adaptive adjustment reduces the need
Application Domain	General optimization problems	General optimization problems	General optimization problems	Stochastic games	Multi-agent reinforcement learning

Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Methodology

- Single-Critic MADDPG

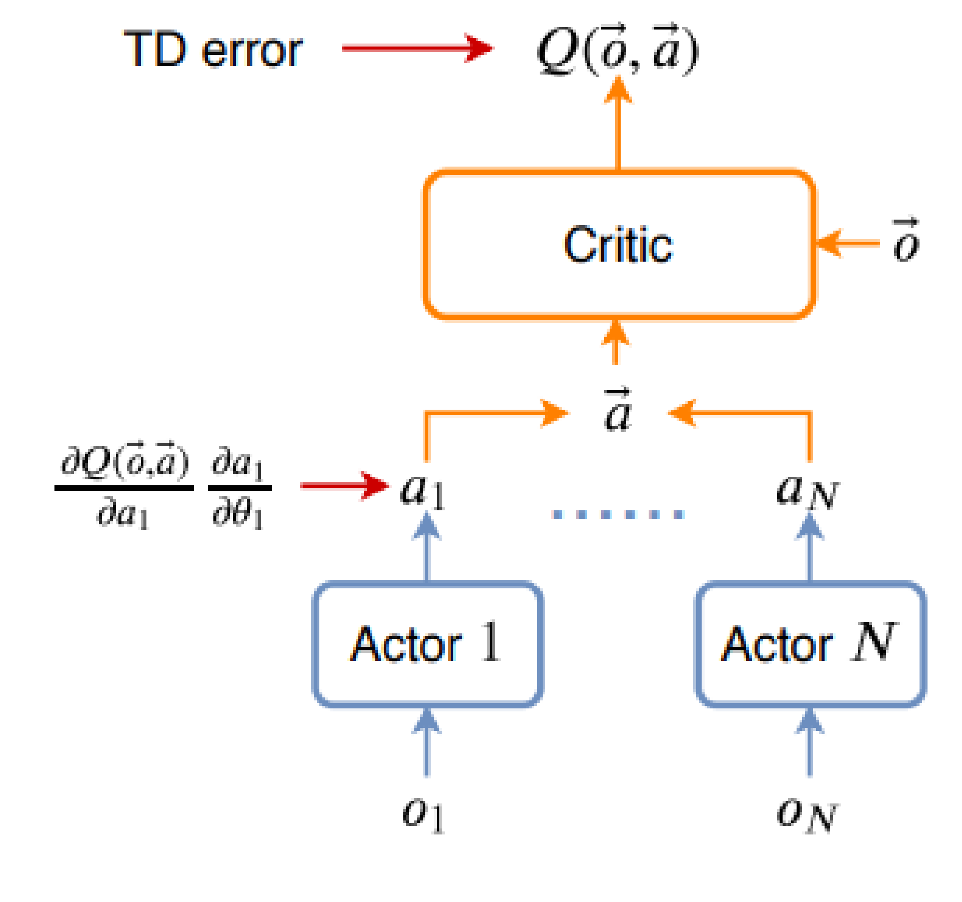


Figure 1: MADDPG

Methodology

- Actor i learning rate: l_{a_i}
- Critic learning rate: l_c
- Adaptive $\overrightarrow{l_a}$ Direction

$$\vec{l_a} = \alpha \vec{l_a} + (1-\alpha)\eta \frac{\partial Q}{\partial \theta} \frac{\vec{\partial} Q}{\partial \theta}^T / || \frac{\partial Q}{\partial \theta} \frac{\vec{\partial} Q}{\partial \theta}^T || \\ \vec{l_a} = \vec{l_a} \frac{\eta}{||\vec{l_a}||}, \tag{1}$$

- Adaptive l_c and $||\overrightarrow{l_a}||$

$$l_c = \alpha l_c + (1-\alpha) l \cdot \mathrm{clip}(| \frac{\partial \delta}{\partial \phi} \frac{\partial Q}{\partial \phi}^T | / m, \epsilon, 1-\epsilon) \\ \eta = l - l_c. \tag{2}$$

Methodology

- First-Order Approximation:
  - The actor i's contribution to ΔQ is only related to the change of action i, without capturing the joint effect with other agents' updates.
  - When agents are strongly correlated, summing up individual updates from each actor doesn't adequately capture the increase in Q-value.
- Second-Order Approximation:
  - The second-order approximation allows the model to account for the interactions between agents' behaviors, enhancing prediction accuracy .

Methodology

- Algorithm: AdaMa on MADDPG

Algorithm 1 AdaMa on MADDPG

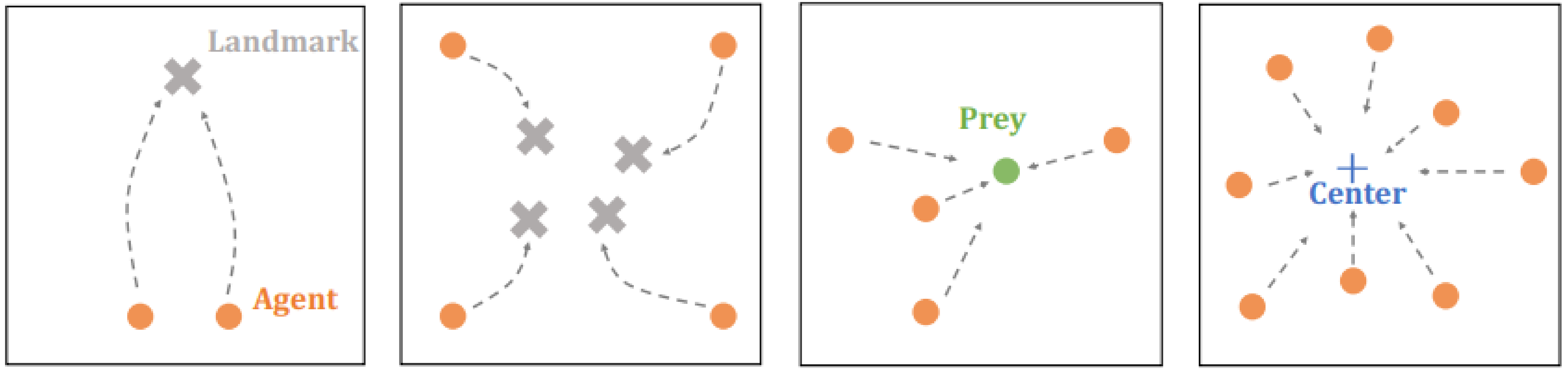
```
Initialize critic network φ, actor networks θ_i, target networks,
  and the replay buffer 𝒟.
Initialize the learning rates l_c and l⃗_a.
for episode = 1, ..., ℳ do
    for t = 1, ..., 𝒯 do
        Select action a_t^i = π_i(o_t^i) + 𝒩_t^i for each agent i
        Execute action a_t^i, obtain reward r_t, and get new observa-
          tion o_{t+1}^i for each agent i
        Store transition (o⃗_t, a⃗_t, r_t, o⃗_{t+1}) in 𝒟
    end for
    Sample a random minibatch of transitions from 𝒟
    Adjust l_c and ||l⃗_a|| by (2).
    Adjust l⃗_a by (1) (first order) or (3) (second order).
    Update the critic φ by φ = φ − l_c ∂δ/∂φ.
    Update the actor θ_i by θ_i = θ_i + l_{a_i} (∂Q(o⃗,a⃗)/∂a_i)(∂a_i/∂θ_i) for each agent.
    Update the target networks.
end for
```

Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Experiments

- Four cooperative scenarios based on Multi-Agent Particle Environment (MPE)

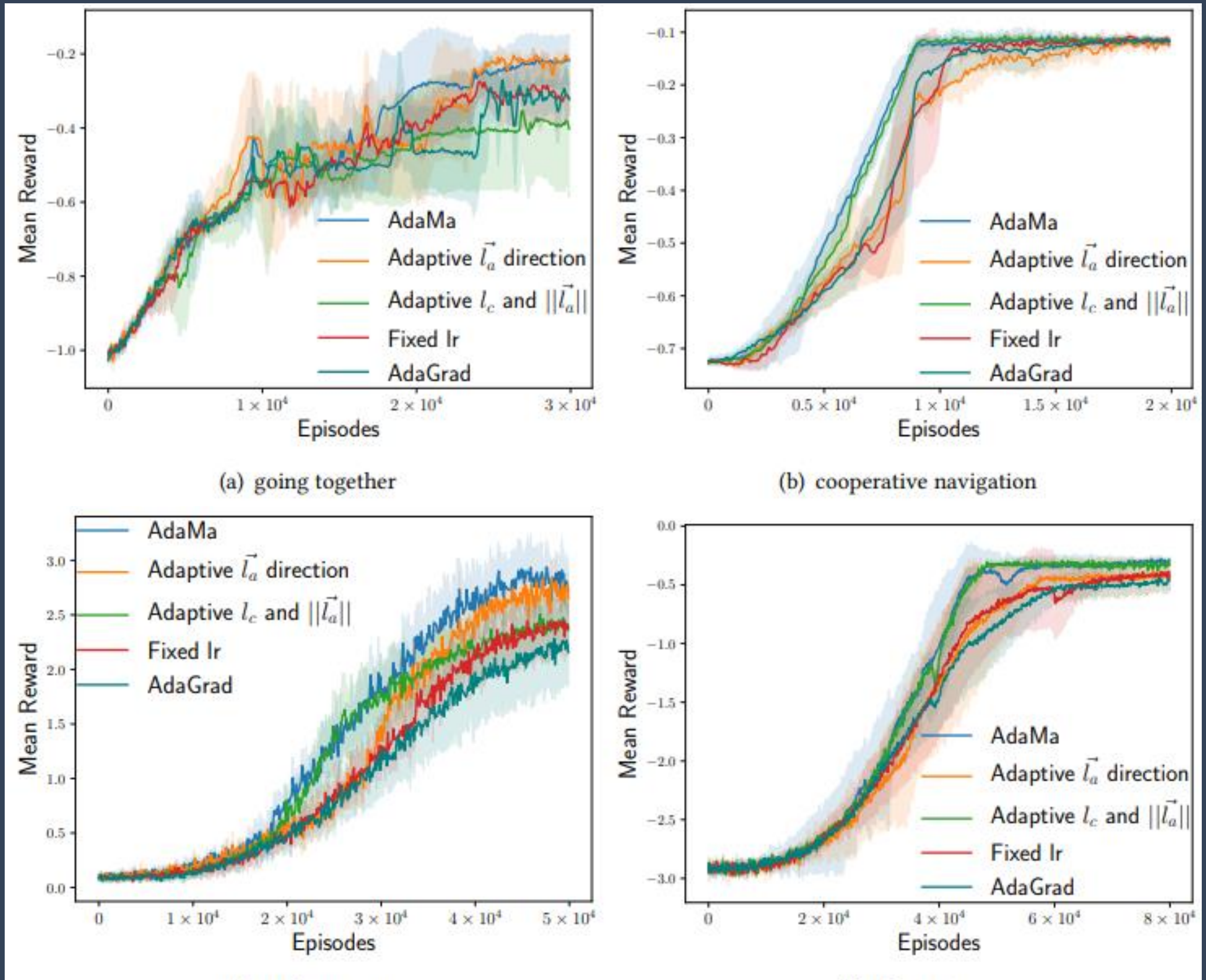


(a) going together (b) cooperative navigation (c) predator-prey (d) clustering

$-0.5(d_i + d_j) - d_{ij}$ $-\max_i(d_i)$ +1 when touched $-\sum d_i$

Experiments

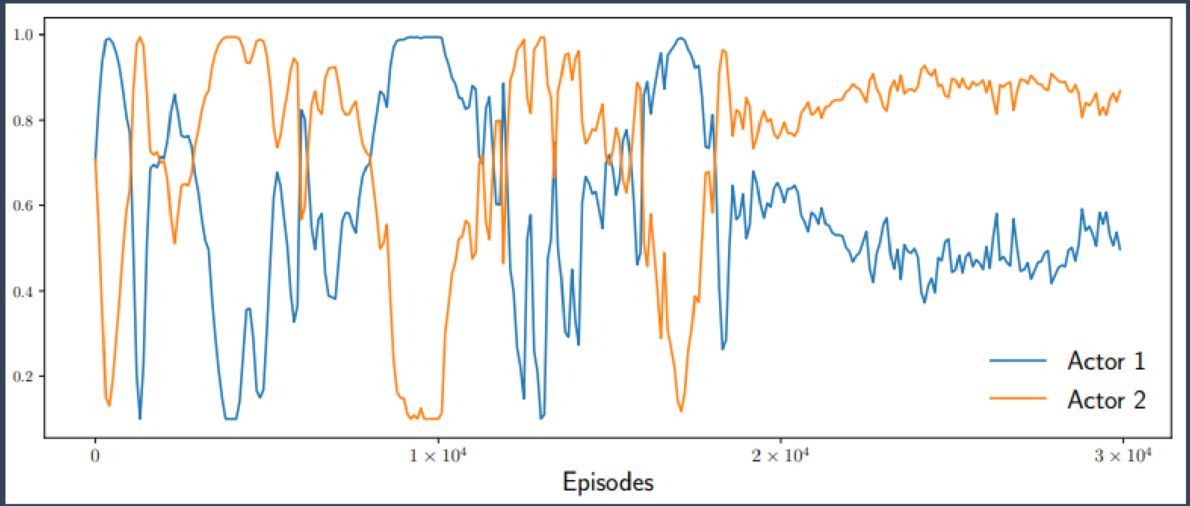
- Learning curves in the four scenarios.



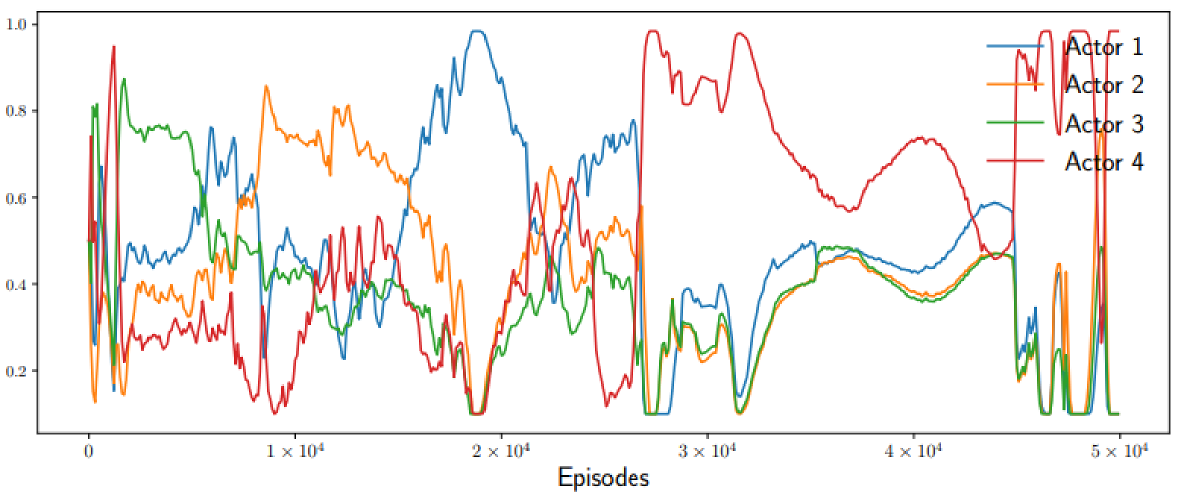
(a) going together
(b) cooperative navigation
(c) predator-prey
(d) clustering

Experiments

- Normalized actors' learning rates during the training in going together.

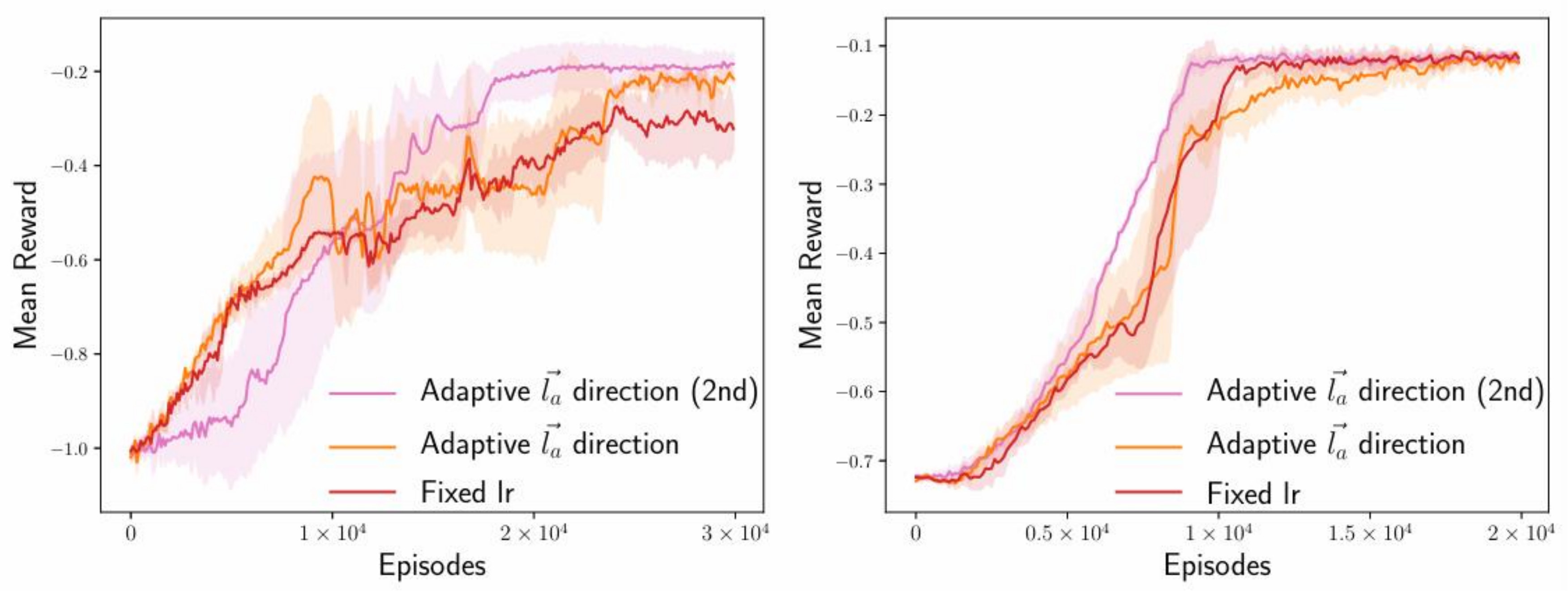


- Normalized actors' learning rates during the training in predator-prey.



Experiments

- Learning curves with the second-order approximation.

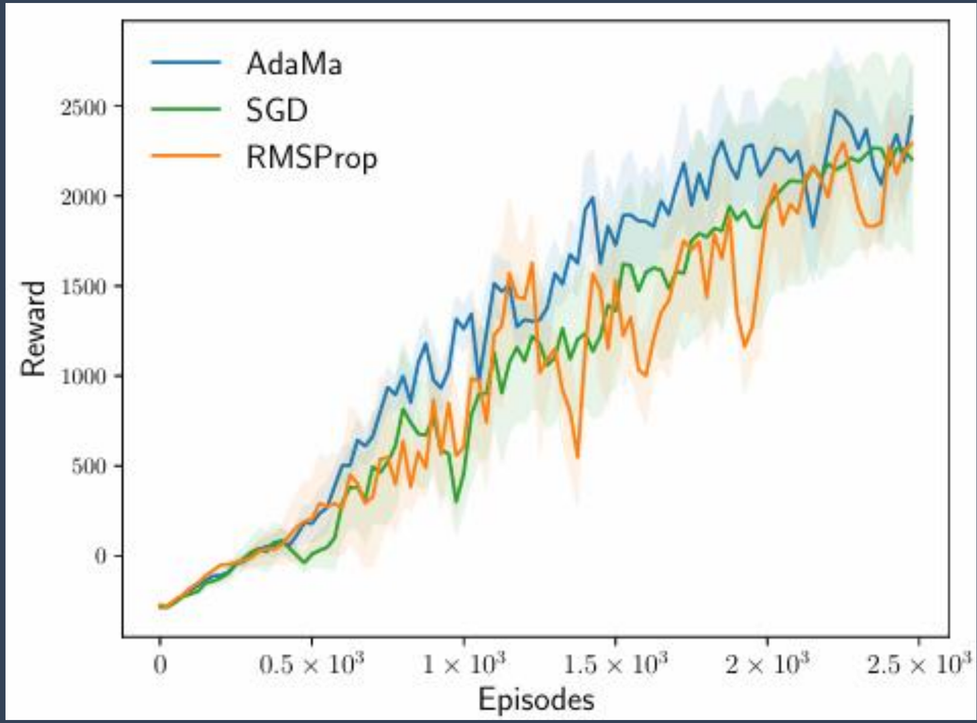


(a) going together (b) cooperative navigation

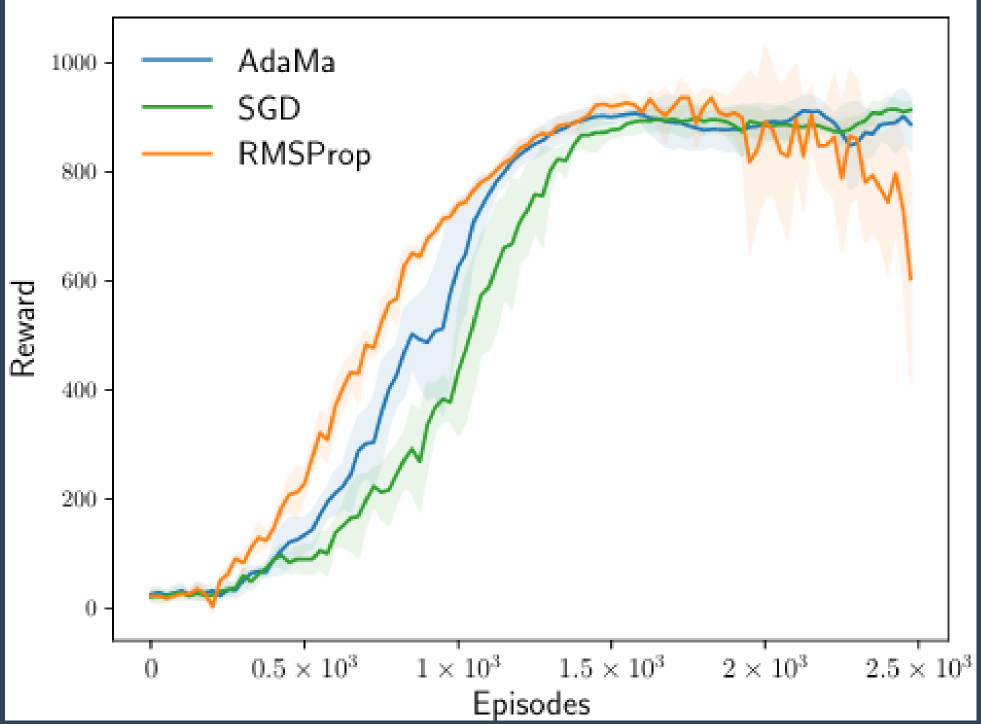
Experiments

- Learning curves of AdaMa on multi-agent mujoco.

HalfCheetah-MADDPG



Ant-MADDPG



Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

Conclusion

- Pros:
  - AdaMa adaptively update the learning rate in multi-agent environments, effectively accelerating learning.
  - AdaMa can be applied to various multi-agent scenarios with a single critic.
- Cons:
  - In some environments, using AdaMa may result in less effective learning in the early stages compared to other methods with adaptive learning rates.

Conclusion

- Future work:
  - The current work on AdaMa does not cover environments with multiple critics in multi-agent systems, therefore, AdaMa can be modified in the future to suit environments with multiple critics in multi-agent systems.

Q & A