# Meta Learning for Meta-Surface: A Fast Beamforming Method for RIS-Assisted Communications Adapting to Dynamic Environments

Qinpei Luo and Boya Di
School of Electronics Engineering and Computer Science, Peking University, Beijing, China

## Introduction

- Reflective Intelligent Surface (RIS): A promising technique to enhance the capacity of wireless networks by its ability of desirable signal reflection.
- Two Main Challenges for RIS in real world:
  - Numerous RIS elements---Large Solution Space;
  - Dynamic Environments---Out-of-date Solution.
- Our proposed Method---MC-DDPG
  - A meta-critic based Reinforcement Learning framework that recognizes the environment change and automatically perform the self-updating of model when environment varies.
  - A stochastic Explore and Reload procedure to alleviate the high-dimensional action space issue.

## System Model

- A downlink multi-user MISO wireless communication system shown as (Fig. 1): One BS, four users, one RIS reflecting the signal from the BS.

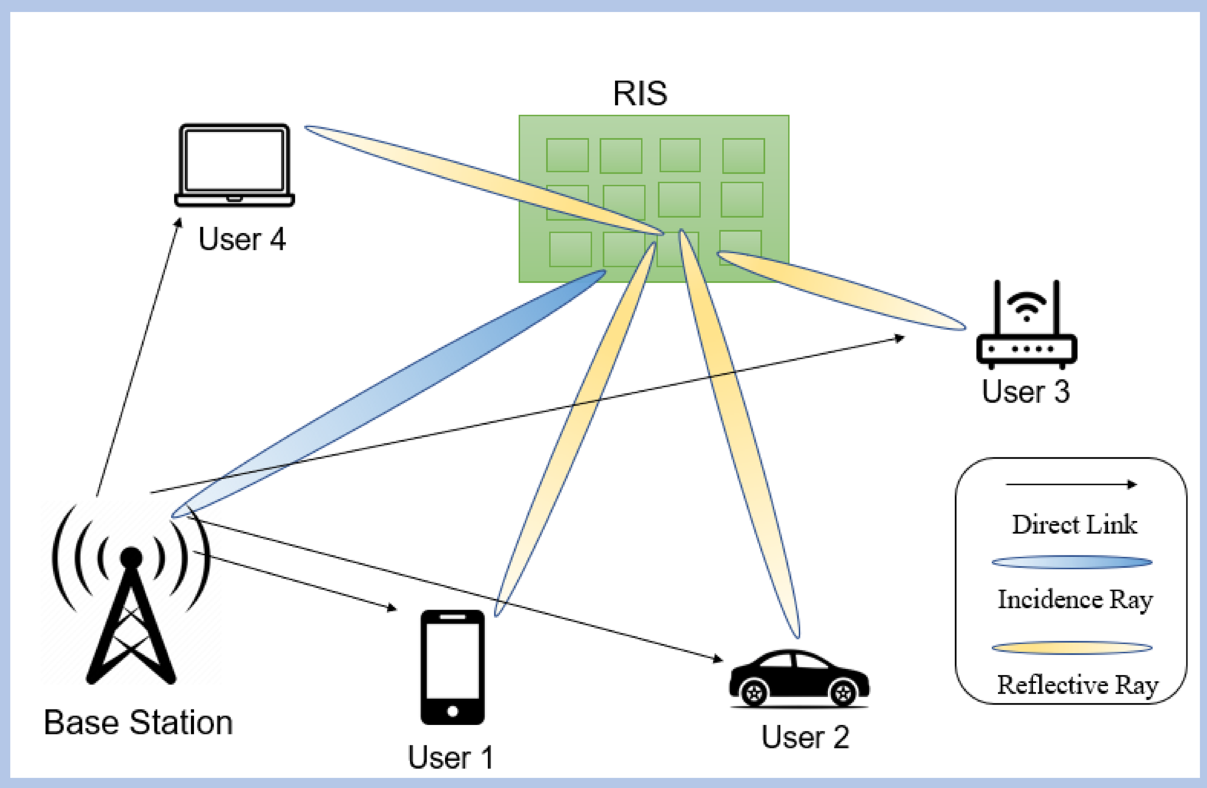


**Fig. 1: System model**

- The channel between BS can be divided into two components: direct channel $\boldsymbol{H}_{BU}$, and channel from BS to RIS and RIS to users, denoted by $\boldsymbol{H}_{BR}$ and $\boldsymbol{H}_{RU}$. The equivalent end-to-end channel can be expressed by
$$\boldsymbol{H}_k = \boldsymbol{H}_{RU,k}\boldsymbol{\Theta}\boldsymbol{H}_{BR} + \boldsymbol{H}_{BU,k}$$
- $\boldsymbol{\Theta} = \mathrm{diag}([e^{j\theta_1}, e^{j\theta_2}, \dots, e^{j\theta_N}])$ is the phase shift configuration of RIS elements.

## Problem and Markov Decision Process (MDP) Formulation

- We consider the sum-rate maximization problem respect to the phase shift configuration of RIS.

$$\gamma_{k,t} = \frac{\left|(\boldsymbol{H}_{RU,k}\boldsymbol{\Theta}\boldsymbol{H}_{BR}+\boldsymbol{H}_{BU,k})\mathrm{v}_{k,t}\mathrm{s}_k\right|^2}{\left|(\boldsymbol{H}_{RU,k}\boldsymbol{\Theta}\boldsymbol{H}_{BR}+\boldsymbol{H}_{BU,k})\sum_{j=1,\neq k}^{K} V_{j,t}s_j\right|^2+n_{k,t}^2}, R_{k,t} = \left|\Delta T log\left(1+\gamma_{k,t}\right)\right|.$$

$$\boldsymbol{max}_{\boldsymbol{\Theta}_{\mathbf{t}}} \sum_{k=1}^{K}\sum_{t=1}^{T} R_{k,t}$$

- $\boldsymbol{V}_{k,t}$ represents the digital beamforming vector from the BS to $j$-th user, which is given by a fixed beamforming scheme ZF or MMSE, $s_k$ denotes the symbol sent to user $k$ from BS. $n_{k,t}$ denotes the gaussian noise which follows $N(0, \sigma_{k,t}^2)$.
- Given the time-varying characteristics of channels, we then reformulate it as a MDP consisting of the following components
  - **Action:** $\boldsymbol{a_t} = \boldsymbol{\Theta}_{\mathbf{t}}, \forall\theta_t \in \boldsymbol{\Theta}_{\mathbf{t}}, \theta_t \in (-\frac{\pi}{2}, \frac{\pi}{2})$.
  - **State:** $\boldsymbol{s_t} = \{\boldsymbol{H}_t, \boldsymbol{\Theta}_{t-1}\}, \boldsymbol{H}_t = \boldsymbol{H}_{RU,t}\boldsymbol{\Theta}_{\mathbf{t}}\boldsymbol{H}_{BR,t} + \boldsymbol{H}_{BU}$.
  - **Reward:** $\boldsymbol{r_t} = \boldsymbol{\eta}\sum_{k=1}^{K} R_{k,t}$, where $\eta$ is a constant coefficient.

## Motivation of Meta Learning

- In real-world dynamic settings, we face two challenges: 1) Difficulty to obtain datasets; 2) Out-of-date solution.
- Our aim: A model that can "learning to learn". It can automatically identify the task and update its model quickly to converge with fewer data collected, as long as it is pre-trained well on multiple tasks.

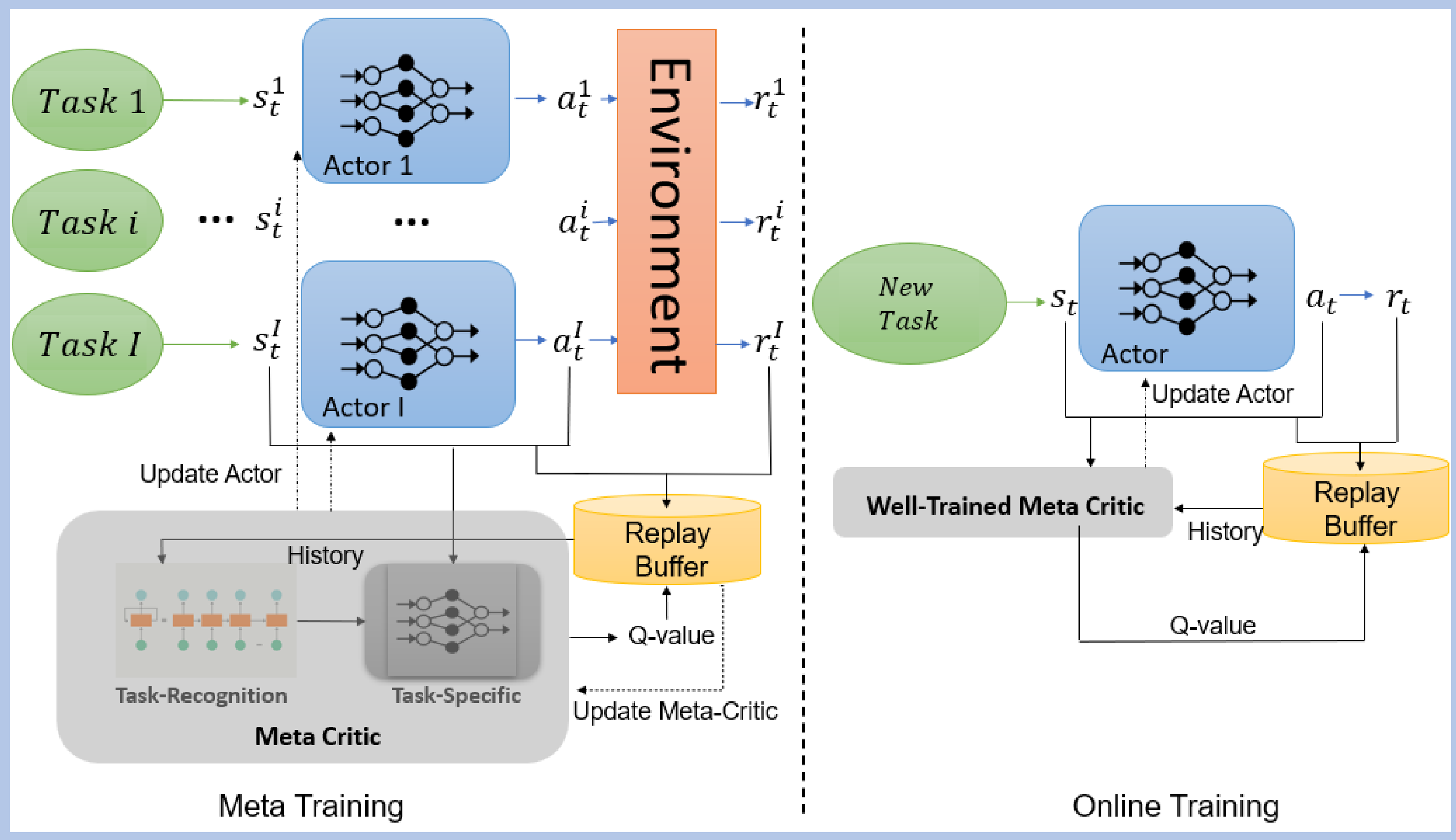


**Fig. 2: The Proposed MC-DDPG Algorithm**

## Algorithm Description

- Task: A process of the BS maximizing the aggregate sum rates of all users. For different tasks, the channel states and locations of users are different.
- Meta Learning Phase:
  - For each task, the current state is fed to the actor to generate an action.
  - Then it operates the action and get a reward from the environment.
  - The transition tuple of state, action and reward will be stored in the replay buffer to form the history.
  - The meta-critic collects history information and current state-action pair to output the task-specific Q-value, which is used to update the actor.
  - The meta-critic is also updated by the trajectories in replay buffer.
- Online Learning Phase:
  - For a newly-coming task, the update of actor is the same as in the Meta Learning Phase.
  - The well-trained Meta Critic is kept static.

## Simulations and Conclusions

- The proposed algorithm is compared to two benchmarks:
  - Zero-Force Exhausting
  - Twin delayed deep deterministic policy gradient (TD3)

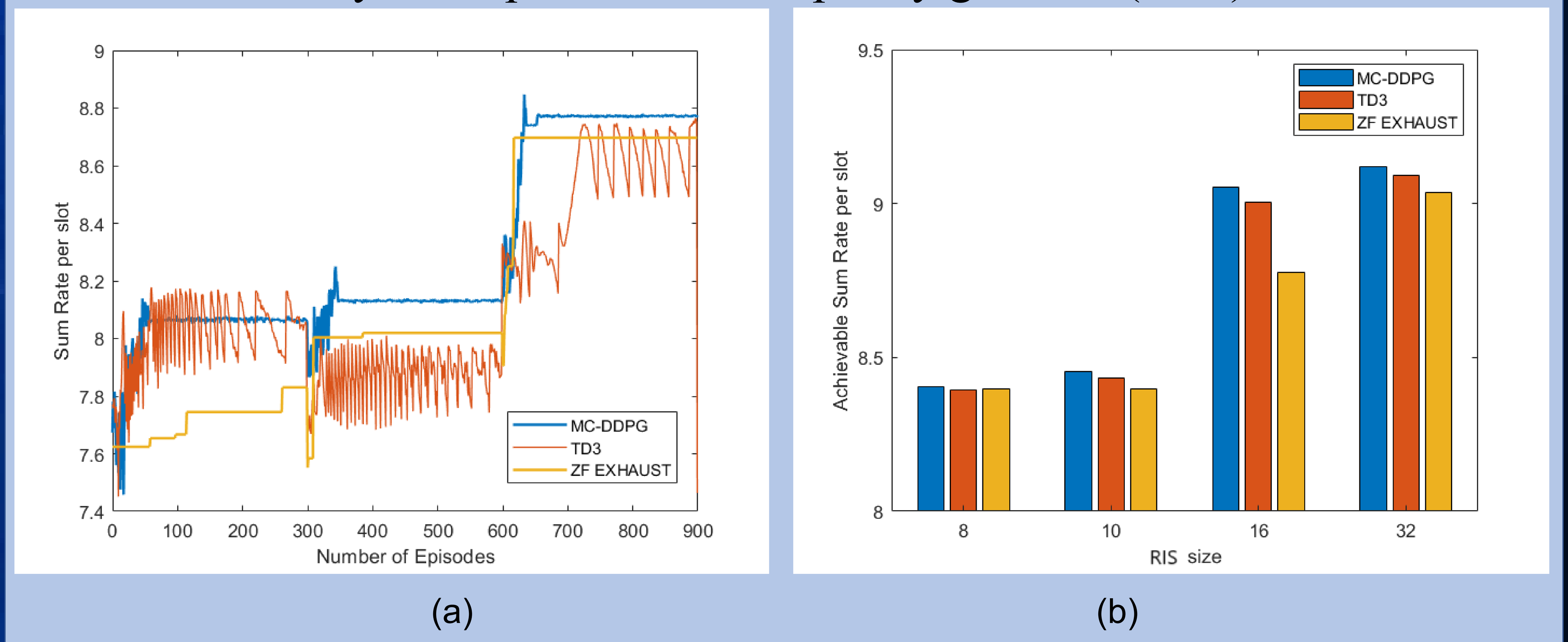


**Fig. 3: Simulation results: a) Sum rate performance with respect to varying users' locations; b) Achievable sum rate vs. the number of RIS elements**

- Fig. 3(a) shows the performance of the proposed scheme when mobile users move rapidly. The proposed MC-DDPG can rapidly converge to a higher sum rate compared to the benchmarks.
- As the MC-DDPG can converge within 100 episodes (1ms) and we set each user's position changes 0.01m each episode, we remark that it can support the user mobility at a minimum speed of 36 km/h.
- Fig. 3(b) shows the sum rate varying with the number of RIS elements N. We observe that the proposed MC-DDPG converges to a higher sum rate as N increases, which shows its capability of supporting large-scale RIS.