Reinforcement Learning based Efficient Combinatorial Optimization of Binary RIS patterns

Sung Woo Cho, Jina Lee, Kae Won Choi*
Sungkyunkwan Univ., Sungkyunkwan Univ., *Sungkyunkwan Univ.
luckycho@g.skku.edu, ginalee0421@skku.edu, *kaewonchoi@skku.edu

강화학습 기반 이진 RIS 패턴의 효율적 조합최적화

조성우, 이지나, 최계원*
성균관대학교, 성균관대학교, *성균관대학교

Abstract

We present a PPO-based reinforcement learning approach to solve the vast combinatorial optimization challenge of RIS pattern, efficiently navigating the binary configuration space to achieve precise directional control in near-field environments.

Ⅰ. Introduction

Reconfigurable Intelligent Surfaces (RIS) have emerged as a key technology for next-generation wireless communications by controlling electromagnetic wave propagation with minimal power consumption. The critical challenge in RIS optimization is determining the optimal configuration from 2^{N^2} possibilities for an N×N array, creating a search space intractable for traditional methods when applied to practical arrays like 16×16 elements. We address this challenge with a Proximal Policy Optimization (PPO) framework that combines a dual-paradigm exploration strategy, a composite reward function with MSE and contrast metrics, and a CNN-based policy network for hierarchical feature extraction. Experimental results demonstrate superior beamforming performance compared to conventional approaches, with significant implications for practical RIS deployment in next-generation communication systems.

Ⅱ. Method

A. RIS System Configuration and Near-field Modeling

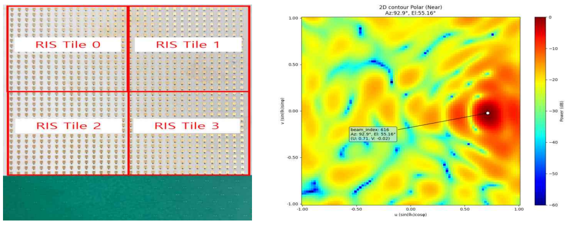

Fig 1. RIS hardware & nearfield beam pattern

Our prior work has implemented a 32×32 1-bit unit cell RIS operating at 26.5–29.5 GHz. The beamforming algorithm employs a codebook approach with 1,024 predefined beam directions in the U-V domain, incorporating near-field considerations including distance variations and spherical wavefront curvature effects.[1]

$$AF_{near} = \exp(-jk(-r + p \cdot u + x))$$

For near-field modeling, we define r as the reference distance from RIS center to receiver, P as the RIS element position vector, and u as the direction vector. The phase difference calculation incorporates the spherical wavefront curvature effect, which we approximate using Taylor series expansion:

$$x = \frac{|p|^2 - (p \cdot \hat{u})^2}{2r} = \omega_{rx} \cdot |p|^2, \text{ where } \quad \omega_{rx} = \frac{RIS_{cell_{distance}}}{2 \cdot Rx - RIS_{cell_{distance}}}$$

To validate our PPO-based reinforcement learning framework, we conducted experiments using a 16×16 RIS tile array with binary states (0/1), creating a vast 2^{256} configuration space. The target radiation pattern aimed to optimize beamforming toward a specific near-field direction (r=0.05m, azimuth=92.9°, elevation=55.16°).

B. Reinforcement Learning Architecture

The proposed RIS pattern optimization architecture integrates various components to enable end-to-end learning. The RIS pattern is represented as an N×N binary matrix, with N=16 in this experiment. The radiation pattern simulation function R takes the binary pattern P as input and outputs a 2D radiation pattern. Specifically, M=91, and the radiation pattern is represented on a 91×91 grid in dB.

$$F = R(P) \in R^{M \times M}$$

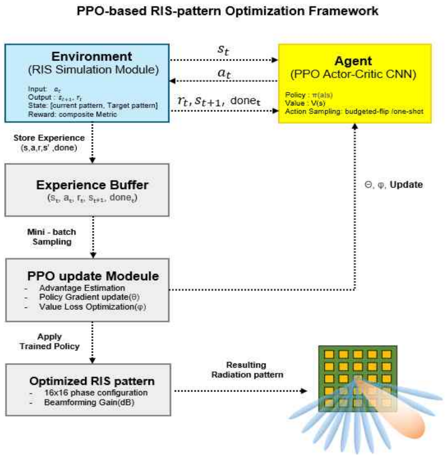

Fig 2. PPO-based RIS-pattern Optimization Framework

The environment evaluates radiation patterns from binary configurations and provides performance-based rewards to the agent.

$$s_t=[F_t,F^{target}]\in R^{2\times M\times M}\quad ,\ s_t\in S$$

The state space(S) comprises two normalized 91×91 dB-scale radiation patterns (current and target pattern) as a two-channel input, enabling direct pattern comparison. This design enables the agent to compare the current and target patterns, facilitating efficient learning.

$$A=\{0,1\}^{N^2}\ ,\ a_t\in\{0,1\}^{N^2}$$

The action space(A) features two modes: "one-shot" (setting all 256 elements simultaneously) and "budgeted-flip" (sequentially flipping individual elements with optional termination), enabling flexible optimization approaches.

$$A=\{1,2,...,N^2,N^2+1\},\ a_t\in\{1,...,N^2,N^2+1=stop\}$$

The transition function(T) defines the state update rule given the current state and selected action:

$$F_{t+1}=R(P_{t+1})\ ,\ s_{t+1}=[F_{t+1},F^{target}]$$

The state(s_{t+1}) observation combines current and target radiation patterns for direct comparison. The reward function outputs a scalar value based on weighted performance metrics, primarily using MSE between current and target patterns.

$$MSE=\frac{1}{M^2}=\sum_{i=1}^{M}\sum_{j=1}^{M}(F_t[i,j]-F^{target}[i,j])^2$$

The reward function combines target intensity rewards with contrast metrics to differentiate focal and background areas. Our PPO implementation uses advantage estimation and objective clipping for stability, with a temperature parameter(τ_t) controlling the exploration-exploitation balance during training.

The Actor-Critic network employs a CNN encoder that downsamples the 91×91×2 input (current and target patterns(s_t)) to extract a 384-dimensional(d) feature vector(z_t).

$$z_t=f_{\theta_{enc}}(s_t)\in R^d$$

The Actor network transforms the feature vector into a categorical distribution over 257 actions (256 element positions plus one stop action) through multiple MLP layers(f_{θ_π}), while the Critic network estimates expected future rewards from the same features. This shared-encoder design enables efficient processing of high-dimensional radiation patterns while maintaining stable policy learning.

$$l_t=f_{\theta_\pi}(z_t)\in R^{N^2+1}\ ,\ \pi_{\theta_\pi}(a_t\mid s_t)=Categorical(\frac{l_t}{\tau_t})$$

Ⅲ. Conclusion

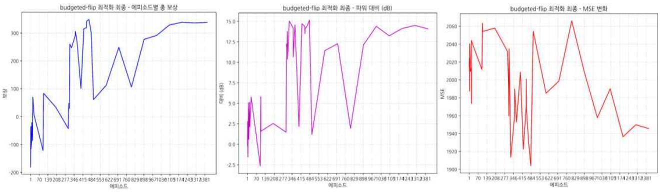

Fig 3. Performance Metric During Reinforcement Learning

Our budgeted-flip approach demonstrated remarkable learning progression from initial low rewards to significant improvement around episodes 200-300, with convergence and early termination at episode 1,481. Key metrics improved consistently: MSE decreased from above 2060 to 1930.27, while power contrast reached 12.95dB, confirming efficient convergence within the vast RIS configuration space.

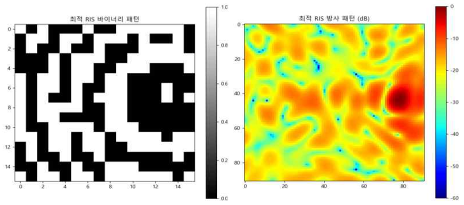

Fig 4. Optimized Binary Pattern and Its Radiation Pattern

Fig. 4 shows the optimal binary pattern derived through reinforcement learning and its resulting radiation pattern. The simulation shows a directional beam precisely at the target position, with concentrated main lobe energy and minimal sidelobes. These results confirm our PPO-based approach effectively navigates the vast 2^{256} configuration space to achieve precise beamforming in challenging near-field conditions.

ACKNOWLEDGMENT

This work was supported by the BK21 FOUR Project.

REFERENCES

[1] Cho, S. W., Lee, J. N., & Choi, K. W. (2025). Implementation of Reconfigurable Intelligent Surface and Experimental Validation of Codebook-Based RIS Beamscanning in mmWave. 한국통신학회 학술대회논문집, 626-627.