# 다중 안테나 네트워크에서의 강화학습 기반 무선 자원 할당에 관한 연구

김다윗, 김도엽, 김경원, 이장원
연세대학교
{greenwings, danny.doyup.kim, kyeongwon.kim, jangwon}@yonsei.ac.kr

# Reinforcement Learning-Based Radio Resource Allocation for Multi-Antenna Networks

David Kim, Do-Yup Kim, Kyeong-Won Kim, Jang-Won Lee
Yonsei Univ.

요 약

본 논문은 다중 안테나 네트워크에서 사용자의 서비스 품질을 만족시키며 높은 네트워크 성능을 제공하기 위해 시간 및 주파수 자원 할당에 관한 문제를 다루고, 해당 문제를 강화학습 기반 알고리즘을 사용하여 해결한다. 제안 알고리즘은 네트워크 환경이 바뀌어도 사용자의 서비스 품질을 보장하면서 효율적으로 자원 할당 정책을 학습함을 모의실험을 통해 보인다.

## Ⅰ. 서 론

무선 자원 할당은 통신 네트워크에서 중요한 분야 중 하나로, 기존에는 최적화 이론과 게임 이론 등의 수학적 기반의 다양한 접근법이 제시되었다[1]. 하지만 통신 네트워크가 점차 거대해지고 복잡해지면서 기존의 수학적 이론을 통한 접근이 어려워지고 있어, 최근 기계 학습 기반의 무선 자원 할당에 관한 연구들이 진행되고 있다[2].

한편, 미래 통신 시스템은 base station (BS)의 안테나 수가 더욱 많아지는 방향으로 발전하고 있다[3]. 대부분의 다중 안테나 네트워크 연구는 링크레벨 관점에서 물리계층 기술에 초점이 맞추어 진행되었고, 시스템레벨에서 사용자 서비스 품질을 보장하며 높은 성능을 위한 무선 자원 할당에 관한 연구는 미미한 실정이다.

본 논문에서는 단일 안테나만을 고려했던 선행 연구 [2]를 확장하여 다중 안테나를 가지는 BS와 실제 환경에 더욱 유사한 시스템을 고려한다. 해당 시스템에서 다중 안테나를 가지는 BS가 실시간으로 변화하는 네트워크 환경에서 사용자의 서비스 품질을 보장하며 높은 전송률을 제공할 수 있는 무선 자원 할당 문제를 다룬다. 강화학습 알고리즘을 이용하여 해당 문제를 해결하는 기법을 제시하고, 모의실험을 통해 성능을 보인다.

## Ⅱ. 시스템 모델 및 문제 형성

본 논문에서는 time division multiple access (TDMA)로 동작하는 단일 셀 시스템을 고려한다. 시스템 내 총 $U$명의 사용자가 분포된 상황에서 $N_t$개의 송신 안테나를 갖는 BS가 매 시간-주파수 자원마다 1명의 사용자를 선정하여 무선 자원을 할당하는 시나리오를 고려한다. 본 논문에서는 사용자와 time slot 인덱스를 각각 $u \in \{1,2.\cdots,U\}$ 및 $n \in \{0,1,2,\ldots\}$으로 표기한다.

채널 모델은 extended Saleh-Valenzuela 모델을 고려하여 사용자 $u$와 BS 사이의 채널 $\boldsymbol{h}_u$를 다음과 같이 정의한다[4].

$$\boldsymbol{h}_u = \sqrt{N_t}\,\alpha \mathbf{a}_{\mathrm{t}}^{\mathrm{H}}(\theta^{\mathrm{t}}). \quad (1)$$

위 식에서 $\alpha$, $\mathbf{a}_t$, $\theta^t \in [0,2\pi)$는 채널 gain, BS의 안테나 steering 벡터, 방위각을 각각 나타낸다.

시간-주파수 자원 관리 문제에서 시스템 상태를 나타내는 state와 자원 할당을 나타내는 action을 다음과 같이 정의한다. time slot $n$에서의 state를 $s(n) \in \boldsymbol{S}$로, action을 $a(n) \in \boldsymbol{A}$로 표기하며 state에 따라 action을 선택하는 정책을 $\pi : \boldsymbol{S} \rightarrow \boldsymbol{A}$로 나타낸다. 각 사용자 $u$의 순시 전송률을 $r_u(s(n),a(n))$로 표기하고, discount factor $\gamma \in [0,1]$를 도입하여 본 연구에서는 다음과 같은 자원 관리 문제를 해결하고자 한다.

$$\underset{\pi \in \boldsymbol{\Pi}}{\text{maximize}}\ \boldsymbol{E}\left[\sum_{n=0}^{\infty}(\gamma)^n \sum_{u=1}^{U} r_u(s(n),\pi(s(n)))\right] \quad (2.1)$$

$$\text{subect to}\ \mathbf{E}\left[\sum_{n=0}^{\infty}(\gamma)^n r_u(s(n),\pi(s(n)))\right] \geq \bar{\delta}, \forall u, \quad (2.2)$$

위 문제에서 $\bar{\delta}_u$는 사용자 $u$의 discounted 최소 요구 전송률을 나타낸다. 위의 constrained Markov decision process (CMDP) 문제는 제약식으로 인해 다루기 어려우므로 Lagrangian 기법을 이용하여 unconstrained Markov decision process (UMDP)로 재구성한다. Lagrangian cost 함수 $c^{\boldsymbol{\mu}}(s,a)$를 사용하여 다음과 같은 UMDP 문제를 재정의한다.

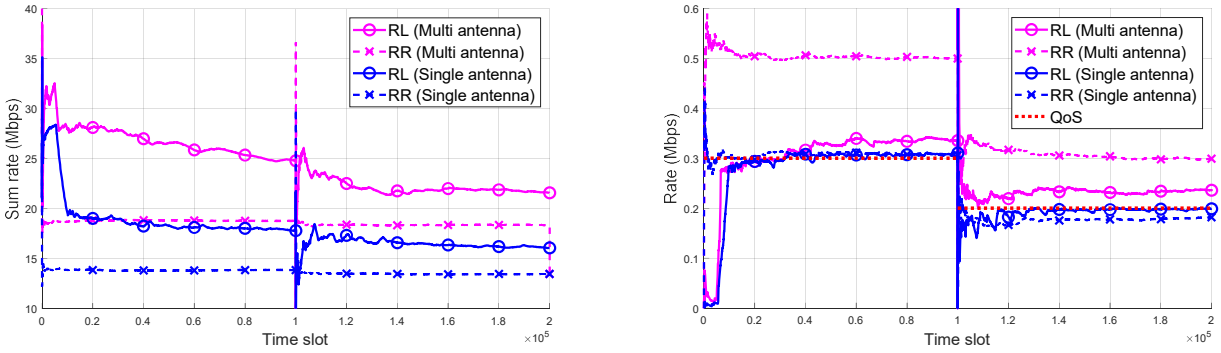
$$\underset{\pi \in \boldsymbol{\Pi}}{\min}\underset{\boldsymbol{\mu} \geq 0}{\max}\ \boldsymbol{E}\left[\sum_{n=0}^{\infty}(\gamma)^n C^{\boldsymbol{\mu}}(s(n),\pi(s(n)))\middle| s^0 = s\right] - \boldsymbol{\mu}^T\boldsymbol{\delta}. \quad (3)$$

위의 UMDP 문제는 최적 Lagrangian multiplier와 최적 정책을 모두 구하여 해결할 수 있다.

## Ⅲ. 문제 해결

네트워크 환경은 실시간으로 변화하기 때문에 이를 적응적으로 학습하여 자원을 할당하는 알고리즘이 요구된다. 이 장에서는 시스템이 사용자 수와 사용자들의 채널 정보가 변화하는 환경에서 효과적인 자원 할당을 수행할 수 있도록 새로운 state 및 action 영역을 제안한다.

Time slot $n$에서 사용자 $u$의 채널 gain $x_u(n)$, 방위각 $y_u(n)$, 서비스 품질 만족도 $z_u(n)$을 사용자 $u$에 대한 state 정보로 선정한다. 연속적 정보를 state로 정의하기 위해 각 정보에 대해 적절한 분할(partitioning)이 필요하다. 이때 state는 환경변화에 독립적이며 state의 각 element는 특




(a) 시스템 전체 성능 (b) 사용자 $w$의 전송률

그림 1. 제안 무선 자원 할당 알고리즘의 성능 검증 결과

정 $(x,y,z)$ 상황을 갖는 사용자의 존재 여부를 의미한다.

시스템은 $U$명의 사용자 중 1명의 사용자에게 자원을 할당하므로 각 action은 자원을 할당받을 사용자를 나타내도록 정의한다. 이때 action이 자원을 할당할 특정 $(x,y,z)$ 상황을 선택하는 것으로 정의한다.

본 논문에서는 시스템 내 사용자들의 현재 상태를 고려하여 최적의 자원 할당을 진행하는 알고리즘을 개발한다. 최적 정책 $\pi^{*,\boldsymbol{\mu}^*}$을 얻기 위해 최적 Lagrangian multiplier $\boldsymbol{\mu}^*$에서의 최적 action-value 함수 $Q^{*,\boldsymbol{\mu}^*}$가 필요함. 그러므로 $\boldsymbol{\mu}^*$와 $Q^{*,\boldsymbol{\mu}^*}$를 함께 학습하도록 한다.

최적 action-value 함수를 학습하기 위해 state에 대한 action-value 값을 출력하는 Q-network (Q-Net)을 구성한다. 본 논문에서는 Q-Net의 학습을 위해서 DQL 기법을 활용하며 최적 Lagrangian multiplier의 학습을 위해서 stochastic sub-gradient 알고리즘을 적용한다.

## Ⅳ. 모의실험 결과

본 모의실험에서는 제안 알고리즘이 동적으로 변화하는 네트워크 환경에서 효율적으로 시간-주파수 자원을 할당할 수 있는지 확인하기 위해 사용자 수, 채널 정보, 요구 전송률이 다른 2개의 시나리오에서 각 $10^5$ time slot 동안의 성능을 확인한다. 각 시나리오의 사용자 수와 최소 평균 전송률 요구 조건은 각각 15명, 0.3Mbps와 25명, 0.2Mbps로 설정한다.

제안 알고리즘의 유효성 검증을 위해 round robin (RR) 알고리즘과 시스템 성능을 비교한다. 또한, 채널 gain이 가장 좋지 않은 사용자가 사용자 중에서 요구 전송률을 만족하기 가장 어려우므로 해당 사용자 $w$의 전송률과 요구 전송률을 비교한다.

그림 1에서 제안 알고리즘은 두 시나리오에서 모두 RR 알고리즘 대비 사용자 $w$의 요구 전송률을 만족시키며 높은 시스템 성능을 보인다. 또한, 다중 안테나 시스템을 고려할 경우, 단일 안테나 시스템 대비 높은 성능을 달성할 수 있다. 이는 다중 안테나 시스템에서 안테나 diversity 구현이 가능하기 때문이다. 특히, 시스템 시나리오가 변화하더라도 제안 알고리즘은 이전 시나리오 데이터에서 자원 할당 정책의 특징을 적절히 추출하여 변화된 시나리오에서도 사용 가능한 자원 할당 정책을 적절하게 학습했음을 확인할 수 있다.

## Ⅳ. 결론 및 향후 연구

본 논문에서는 사용자의 평균 요구 전송률을 만족시키며 사용자의 요구 전송률의 합을 최대화를 위한 자원 할당 문제를 해결하는 기법을 제시하였으며, 제안 알고리즘은 시스템 환경이 변화하더라도 사용 가능한 자원 할당 정책을 학습할 수 있음을 모의실험으로 보였다.

다중 안테나 시스템에서는 신호의 세기와 방향을 beamforiming으로 제어해 각 사용자의 공간 자원을 고려하여 사용자 간 간섭을 최소화하며 동시에 다수 사용자에게 자원을 할당할 수 있다. 본 논문에서 정의한 state는 공간 자원에 대한 정보를 포함하기 때문에 각 사용자의 공간 자원과 그에 따른 간섭을 고려한 자원 할당이 가능할 것이다.

이에 따라, 향후 연구에서 action을 $U$명의 사용자 중 $K$명의 사용자에게 자원을 할당하는 것으로 확장할 것이다. Zero-forcing과 같은 beamforming 기술을 사용하여 $K$명의 사용자의 공간 자원에 따른 간섭을 고려하여 action-value 함수를 학습할 수 있을 것이다[5].

## ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019R1A2C2084870).

## 참 고 문 헌

[1] J.-A. Kwon, B.-G. Kim, and J.-W. Lee, “A unified framework for opportunistic fair scheduling in wireless network: A dual approach,” *Wireless Netw.*, vol. 16, pp. 1975-1986, Feb. 2010.

[2] H.-S. Lee, J.-Y. Kim, and J.-W. Lee, “Resource allocation in wireless networks with deep reinforcement learning: A circumstance-independent approach,” *IEEE Syst. J.*, vol 14, no. 2, pp. 3857-3868, Jun. 2020.

[3] W. Ni, and X. Dong, “Hybrid block diagonalization for massive multiuser MIMO systems,” *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201-211, Jan. 2016.

[4] A. A. M. Saleh and R. A. Valenzuela, “A statistical model for indoor multipath propagation,” *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128 - 137, Feb. 1987.

[5] G. Dimic, and N. D. Sidiropoulos, “On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm,” *IEEE Trans. Signal Proc.*, vol. 53, no. 10, pp. 3857-3868, Oct. 2005.