사물인터넷융복합논문지 Vol. 11, No. 5, pp. 189-197, 2025 https://doi.org/10.20465/KIOTS.2025.11.5.189

UAV 기반 RSMA 네트워크에서 사용자 Fairness 향상을 위한 전송 제어

노원종*
한림대학교 소프트웨어학부 교수

Transmission Control for User Fairness in UAV-Assisted RSMA Networks

Wonjong Noh*
Professor, School of Software, Hallym University

요 약 Rate-Splitting Multiple Access (RSMA)와 Unmanned Aerial Vehicles (UAV)는 미래 6G 네트워크에서 연결성과 자원 효율성을 향상시키기 위한 핵심 기술로 주목받고 있다. 본 논문에서는 UAV 기반 RSMA 네트워크에서, UAV의 이동 경로, 빔포밍, 공통 메시지 전송률을 동시에 최적화하여, 사용자들의 최소 전송률을 최대화 함으로서, 사용자들의 공정성 (fairness)를 최대화하는 알고리즘을 제안하였다. 이를 위해, 본 논문은, 마르코프 결정 과정(MDP)를 이용하여 문제를 먼저 정의하고, 이를 해결하는 Proximal Policy Optimization(PPO) 기반 심층 강화학습(DRL) 알고리즘을 제안하였다. 제안하는 PPO 알고리즘은, UAV가 정확한 채널 상태 정보(CSI)에 의존하지 않고도, 동적으로 변하는 환경에서도 효율적으로 간섭 관리 및 자원 할당을 가능하도록 하였다. 시뮬레이션을 통해, 본 논문에서 제안하는 방법이, Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), Trust Region Policy Optimization (TRPO), REINFORCE, Greedy, Random 과 같은 주요 방식들 보다 훨씬 향상된 사용자 최소 전송률을 제공함을 확인하였다. 이를 통해, 제안하는 시스템이 UAV 기반 차세대 무선 네트워크에서 효과적으로 이용될 수 있음을 확인하였다.

주제어 : 딥러닝, 근사 정책 최적, 공평성 최적화, 전송률-분할 다중 접속, 무인 비행체

Abstract Rate-Splitting Multiple Access (RSMA) and Unmanned Aerial Vehicles (UAVs) have emerged as key technologies for enhancing connectivity and resource efficiency in future 6G networks. In this paper, we propose an algorithm that maximizes user fairness (or equivalently maximizes the minimum user data rate) by jointly optimizing the UAV trajectory, beamforming, and common message transmission rate in a UAV-assisted RSMA network. To achieve this, we formulate the problem using a Markov Decision Process (MDP) framework and propose a Proximal Policy Optimization (PPO)-based Deep Reinforcement Learning (DRL) algorithm to solve it. The proposed PPO algorithm enables efficient interference management and resource allocation even in dynamically changing environments, without relying on accurate Channel State Information (CSI). Simulation results demonstrate that the proposed method significantly outperforms existing approaches such as Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), Trust Region Policy Optimization (TRPO), REINFORCE, Greedy, and Random schemes in terms of the minimum user data rate. These results confirm that the proposed system can be effectively utilized in UAV-assisted next-generation wireless networks.

Key Words : Deep reinforcement learning, Fairness Maximization, Proximal policy optimization, Rate-splitting multiple access, Unmanned aerial vehicles

본 논문은 2025년도 한림대학교 교비연구비(HRF-202501-011)에 의하여 연구되었음.
*교신저자 : 노원종(wonjong.noh@hallym.ac.kr)
접수일 2025년 09월 02일 수정일 2025년 10월 02일 심사완료일 2025년 10월 17일

1. 서론

UAV 무선 네트워크는 3차원 공간에서 UAV의 유연한 이동성을 활용하여, 다양한 통신 환경에서 서비스의 품질(QoS)을 크게 향상 시킬 수 있는 시스템으로서, 최근 활발히 연구되어 오고 있다 [1, 2]. 그러나, UAV 무선 네트워크가 다중 사용자 환경에서 효율적으로 이용되기 위해서는, 효율적인 간섭 및 자원관리가 절대적으로 필요하다.

이를 해결하기 위해, 다양한 다중 접속(MA) 기술들이 최근에 많이 연구되고 있다. 예를 들어, 공간 분할 다중 접속(SDMA)은 프리코딩을 통해 공간 분리를 활용하여 여러 사용자에게 동시 전송을 가능하게 한다. 그러나 간섭이 심하거나 이동성이 높은 환경에서는 효과가 제한적이다. 또한, 비직교 다중 접속(NOMA)은 전력 영역에서 사용자를 중첩시켜 스펙트럼 효율을 향상시킬 수 있다. 하지만, 연속 간섭 제거(SIC)로 인해 수신기에 큰 계산 부담을 주고 불완전한 채널 상태 정보(CSI)에서 성능이 크게 저하된다 [3, 4]. 최근에는 SDMA와 NOMA를 포괄하는 보다 일반화된 속도 분할 다중 접속(RSMA)이 제안되고 있다 [5]. RSMA는 모든 사용자를 위한 공통 메시지 부분과 개별 사용자를 위한 전용 메시지 부분으로 메시지를 분할하여, 보다 유연한 간섭 관리를 수행한다. 사용자는 간섭의 일부를 디코딩하고 나머지를 잡음으로 처리할 수 있어서, CSI 불확실성에 대한 견고성이 향상되고 사용자 공정성이 개선됨이 확인되었다 [6]. 다른 연구 [7]는 다양한 무선 시나리오에서 RSMA의 이론적 근거와 실제 적용에 관한 포괄적인 연구를 수행하였으며, 사용자 이질성, 불완전한 CSI 정보, 동적 채널 상태 변화와 같은 네트워크 환경에서, RSMA가 매우 효율적임을 밝혔다. 이러한 장점들 때문에 RSMA는 이동성 및 데이터 전송 채널 변화가 불가피한 UAV 기반 네트워크에서 효율적으로 적용될 수 있음이 입증되었다 [8]-[11].

1.1 관련 연구

최근에, 사용자 공정성 (fairness)을 향상시키기 위한 RSMA 기반 전송 기법에 대한 연구가 많이 진행되어왔다. Dizdar et al. [12]은 사용자 수가 송신 안테나 수를 초과할 때 MIMO 시스템에서 최대-최소 공정성을 달성하는 저복잡도 빔포밍 알고리즘을 제안하였다. 이 방법은 과도한 볼록 최적화를 피함으로써 복잡성과 성능의 균형을 효과적으로 맞추었다. Lee et al. [13]은 제한된 CSI 오류 모델하에서도 견고하게 동작하는 빔포밍 프리코더를 설계하였다. Xu et al. [14]은 유한 길이의 메시지 블록을 가정하에, RSMA 시스템에서 사용자 fairness와 저지연성을 동시에 추구하는 제어 기법을 제안하였다. 그러나 이 연구는 semi-static 또는 low-mobility 환경에 초점을 맞추어 수행되었다.

한편, UAV 기반 네트워크에서 RSMA를 적용하면 이동 환경에서 정보 처리량과 에너지 효율을 개선할 수 있는 것으로 밝혀지고 있다. Jafar et al.[15]은 RSMA 기반 UAV 시스템에서 다운링크 속도 성능에 대한 분석 연구를 제공하였으며, Rician 페이딩 채널에서 NOMA 및 OMA보다 우수함을 보여주었다. Xiao et al.[15]은 연속 볼록 근사(SCA) 기법을 기반으로, 트래픽 인지 에너지 효율적인 RSMA 자원 할당 알고리즘을 제안했다. Feng et al.[16]은 공간 및 물리적 계층을 분리하는 교대 최적화 알고리즘을 사용하여 UAV 이동경로와 RSMA 자원 할당을 동시에 최적화하였다. 그러나, 여기서 제안된 최적 제어 방식은 실시간 또는 고속 이동성 환경에서 계산 비용이 많이 들고 응답성이 떨어진다. Guan et al. [17]은 비상 상황에서 협력적 UAV들이 상호 협력하에 경로 최적화를 수행하는 다중 에이전트 기반 분산 Proximal Policy Optimization (PPO) 프레임워크를 제안하였으며, 이를 통해, UAV가 중앙 제어 없이 지상 사용자에게 적응적으로 서비스를 제공할 수 있도록 했다. Zhan et al. [18]은 Ray 프레임워크를 사용하여 여러 UAV의 분산 제어를 위한 향상된 PPO 기반 알고리즘을 제안하였으며, UAV 임무에서의 향상된 확장성과 성능을 보여주었다.

그러나, 이러한 기존 연구는 대부분 최적 경로 계획을 별도의 제어 문제로 취급하였으며, RSMA 네트워크에서의 빔포밍 및 공통 메시지 전송률 할당등과 연관해서 동시 최적화를 수행하지는 않았다. 또한, 기존 연구에서는 UAV의 빔 제어를 단일 공간 영역으로 제한하는 ULA 안테나를 주로 가정하고 있으며, 방위각과 고도 차원 모두에서 고해상도 3D 프리코딩을 지원하는 URA 안테나 구조를 이용하는 UAV 최적화에 대해서는 많은 연구가 진행되어 있지 않다 [19].

1.2 기여

본 연구의 주요 기여점들은 다음과 같다.

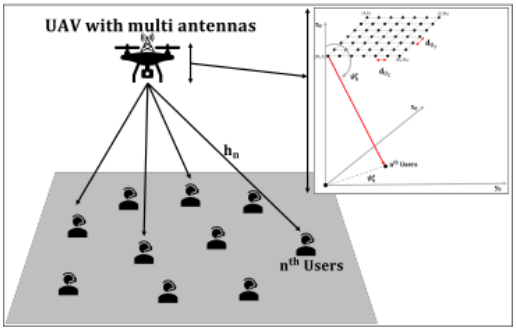
- UVA 기반 RSMA 네트워크에서 빔포밍, 공통 메시지 전송 속도 및 UAV 이동 경로의 동시 최적화를 통하여, 사용자 공정성 (fairness)를 최대로 하는 (사용자의 최소 전송율을 극대화하는) 비볼록

최적화 문제를 설계하였으며, 이에 대한 마르코프 결정 과정(MDP) 문제를 정의하였다.

- 해당 문제를 해결하는, PPO 기반 최적 제어 방안을 제안하였다. 제안하는 방식은, UAV가 이전 채널 상태 정보(CSI)에 의존하지 않으면서도, 복잡하고 동적인 무선 시나리오에서 효율적인 간섭 관리 및 자원 할당이 가능하도록 하였다.
- 시뮬레이션을 통해 본 논문에서 제안하는 방식이 기존 DDPG, SAC, TRPO, REINFORCE, Greedy, Random 과 같은 주요 방식들 보다 훨씬 향상된 사용자 최소 전송률을 제공함을 확인하였으며, 이를 통해, UAV 기반 RSMA 네트워크 제어에, 효과적으로 이용 가능함을 확인하였다.

2. 시스템 모델

본 연구에서는 그림 1에 나타낸 바와 같이 U개의 안테나를 장착한 UAV가 N명의 지상 사용자 위를 비행하는 시나리오를 고려한다. 아래에서, 인덱스 $n=0$은 UAV를 나타내고, $n\in[1,N]$은 지상 사용자를 나타낸다. 시간은 이산 슬롯으로 구성되면, t는 t번째 이산 슬롯을 나타낸다. 각 시간 슬롯은 τ의 시간 크기를 갖는다. UAV는 각 이산 시간 슬롯 마다 m초 동안 이동을 하며, 이동한 후 $|\tau-m|$초 동안 호버링하면서 사용자를 지원한다. UAV의 이동 방향은 각도 $\psi_n(t)\in[0,2\pi]$로 나타내고, 속도는 $V_n(t)\in[0,V_{\max}]$로 나타낸다. 그리고, UAV는 각 시간 슬롯 동안 사용자 위치에 대한 사전 정보를 가지고 있다고 가정한다.



[Fig. 1] System Model

2.1 채널 모델

지상 통신과 달리, UAV 통신은 air-to-ground 채널의 고도 및 고도각에 상당한 영향을 받는다. 따라서, 이러한 UAV 기지국에 적합하고, 강한 LoS를 제공하는 채널 모델링을 위해, 다음과 같은 블록 페이딩 라이시안 페이딩 모델을 가정하였다 [20, 21].

$$\boldsymbol{h}_n = L_n\left(\sqrt{\frac{n^*}{n^*+1}}\boldsymbol{h}_n^{LOS}+\sqrt{\frac{1}{n^*+1}}\boldsymbol{h}_n^{NLOS}\right) \quad (1)$$

여기서 L_n은 대규모 경로 손실을 나타내고, n^*은 라이시안 계수를 나타낸다. 첫째, L_n은 소규모 도시 환경을 가정하여, 다음의 Hata 모델을 가정하였다 [22].

$$L_{un} = 69.55+26.16\log_{10}f+[44.9-6.55\log_{10}h_b]\log_{10}d -13.82\log_{10}h_b-C_H, \quad (2)$$

$$C_H=(1.1\log_{10}f-0.7)h_M+0.8-1.56\log_{10}f \quad (3)$$

여기서 C_H, h_b, h_M, f, d는 각각 안테나 높이 보정계수, 기지국 안테나 높이, 이동국 안테나 높이, 전송 주파수, 기지국과 사용자 간 거리를 나타낸다. 둘째, h_n^{NLOS}은 $CN(0,1)$의 분포를 따르는 것으로 가정하였다. 셋째, h_n^{LOS} 은 다음과 같이 가정하였다 [22].

$$\boldsymbol{h}_n^{LOS}=\boldsymbol{a}_n(\psi_n^e,\psi_n^a)=vec\left(\boldsymbol{a}_{U_1}(\rho)\boldsymbol{a}_{U_2}^T(\zeta)\right) \quad (4)$$

여기서 $a_n(\bullet)\in C^{U\times 1}$은 steering 배열을 나타내고, ψ_n^e은 고도각을 나타내며, ψ_n^a은 방위각을 나타낸다. 또한, vec(·)는 $U_1\times U_2$ 행렬을 벡터화하는 함수이다. $a_{U_1}(\rho)$ 및 $a_{U_2}(\varsigma)$는 다음과 같이 계산된다.

$$\boldsymbol{a}_{U_1}(\rho)=\left[1,e^{j\rho},\ldots,e^{(U_1-1)j\rho}\right]^T \quad (5)$$

$$\boldsymbol{a}_{U_1}(\zeta)=\left[1,e^{j\zeta},\ldots,e^{(U_2-1)j\zeta}\right]^T \quad (6)$$

$$\rho=\frac{2\pi}{\lambda}dU_1cos\psi_n^a sin\psi_n^e=\frac{2\pi}{\lambda}dU_1\frac{x_0-x_n}{d_{0,n}} \quad (7)$$

$$\zeta=\frac{2\pi}{\lambda}dU_2sin\psi_n^a cos\psi_n^e=\frac{2\pi}{\lambda}dU_2\frac{y_0-y_n}{d_{0,n}} \quad (8)$$

이 때, λ는 파장을 나타내고, d_{U_1}과 d_{U_2}는 각각 두 개의 연속된 수직 및 수평 안테나 사이의 거리를 나타낸

다. ρ와 ζ는 고도각과 방위각 벡터를 나타내고, $d_{o,n}$은 UAV와 사용자 사이의 거리를 나타낸다. 여기서, UAV의 위치는 (x_n, y_n)으로 주어진다.

2.2 RSMA 모델

본 연구에서는 1-layer RSMA를 이용한다. RSMA의 세부 설계는 다음과 같다. 사용자를 위한 메시지는 공통 메세지(common message)와 개인 고유 메세지(private message)로 나누어진다. 모든 사용자 간의 공통 메시지는 공통 스트림 s_0에 내장되고, 개인 고유 메시지는 $\{s_n\}$으로 표시된다. 스트림 벡터와 빔포밍 행렬의 정의는 다음과 같다.

$$\boldsymbol{s} = \left[s^c, s_1^p, \ldots, s_N^p\right]^T \in C^{(N+!)\times 1} \quad (9)$$

$$\boldsymbol{P} = \left[\boldsymbol{p}^c, \boldsymbol{p}_1^p, \ldots, \boldsymbol{p}_N^p\right]^T \in C^{(N+!)\times 1} \quad (10)$$

UAV로부터 전송되는 신호는 다음과 같이 표현될 수 있다.

$$\boldsymbol{x} = \boldsymbol{p}^c s^c + \sum_{n\in N} \boldsymbol{p}_n^p s_n^p \quad (11)$$

이 때, 지상의 사용자들은 다음과 같이 신호를 수신한다.

$$y_k = \boldsymbol{h}_n^H \boldsymbol{p}^c s^c + \sum_{j\in N} \boldsymbol{h}_j^H \boldsymbol{p}_j^p s_j^p + n_n \quad (12)$$

여기서, 노이즈는 가산 백색 가우시안 분포 $CN(0,\sigma^2)$를 따른다.

모든 수신자는 처음에 모든 개인 고유 메세지을 간섭으로 처리하면서 공통 메시지를 디코딩한다. 이후, 각 사용자는 다른 개인 고유 메시지를 잡음으로 처리하면서 자신의 개인 고유 메시지를 디코딩한다. 마지막으로, 각 사용자는 디코딩된 공통 메시지와 개인 고유 메시지를 결합한다. 이때, 공통 메시지와 개인 고유 메시지에 대한 신호 대 간섭 및 잡음 비(SINR)는 다음과 같다.

$$\gamma_n^c = \frac{\left|\boldsymbol{h}_n^H \boldsymbol{p}^c\right|^2}{\sum_{j\in N}\left|\boldsymbol{h}_j^H \boldsymbol{p}_j^p\right|^2 + \sigma^2} \quad (13)$$

$$\gamma_n^p = \frac{\left|\boldsymbol{h}_n^H \boldsymbol{p}_n^p\right|^2}{\sum_{j\in N, j\neq n}\left|\boldsymbol{h}_n^H \boldsymbol{p}_j^p\right|^2 + \sigma^2} \quad (14)$$

여기서 γ_n^c과 γ_n^p은 각각 공통 메시지와 개인 고유 메시지에 연관된 사용자 n의 SINR을 나타낸다. 그러면, 달성 가능한 속도 영역은 다음과 같이 표현될 수 있다.

$$R_n^c = \log_2(l + \gamma_n^c) \quad (15)$$

$$R_n^p = \log_2(l + \gamma_n^p) \quad (16)$$

여기서 R_n^c과 R_n^p은 사용자 n이 얻을 수 있는 공통 메시지와 개인 고유 메시지의 전송률을 나타낸다. 반면, 공통 메시지는 모든 사용자가 디코딩해야 하므로, 다음과 같이 모든 사용자의 달성 가능한 전송률 중 가장 작은 전송률로 전송되어야 한다.

$$R_c = \min(R_1^c, \ldots, R_N^c), \quad (17)$$

이를 공통 메시지 전송률 R_c라고 한다. 여기서 R_c는 공통 메시지 전송률 벡터 $c = \{C_1, C_2, \ldots, C_N\}$를 통해 사용자들에게 분할될 수 있다. 여기서 C_n은 각 사용자에게 할당된 공통 메시지 전송률을 나타내며, 이는 다음을 만족한다.

$$\sum_{n\in N} C_n \leq R_c \quad (18)$$

종합하여, 사용자 n이 달성할 수 있는 총 전송률은 공통 메시지 전송률과 개인 고유 메시지 전송률의 합계가 된다.

$$R_n = C_n + R_n^p \quad (19)$$

2.3 문제 설정

본 연구는 사용자간 공정성을 제공하기 위해 지상의 모든 사용자들에게 제공 가능한 최저 속도를 최대화하는 것을 목표로 한다. 이를 위해, 빔포밍 행렬 P, 공통 메시지 전송률 c, UAV 이동 경로 매개변수 $\phi_0(t)$ 및 V_0에 대한 최적화를 고려한다. 즉, 본 연구의 문제는 다음과 같이 표현된다.

$$\max_{\boldsymbol{P}(t), \boldsymbol{c}(t), \boldsymbol{\varphi}_0(t), \mathrm{V}_0(t)} \min_{n\in N} R_n \quad (20)$$

$$s.t. \sum_{n\in N} ||\boldsymbol{p}_n||^2 \leq P_t \quad (21)$$

$$\sum_{n\in N} C_n \leq \min(R_1^c, \ldots, R_N^c) \quad (22)$$

$$C_n \geq 0, \forall n \in N \quad (23)$$

$$0 \leq \varphi_0 \leq 2\pi \quad (24)$$

$$0 \leq V_0 \leq V_{max} \quad (25)$$

여기서, 제약 조건 (21)은 UAV의 모든 사용자를 위한 송신 전력이 $\sum_{n\in N}\|p_n\|^2 \leq P_t$를 만족함을 의미한다. 여기서, P_t는 허용 가능한 최대 송신 전력을 나타낸다. 제약 조건 (22)는 모든 사용자가 공통 메시지를 성공적으로 디코딩할 수 있도록 보장한다. 제약 조건(23)은 전송 속도가 양수 값이 되도록 보장한다. 제약 조건 (24)와 (25)는 UAV 방향각과 속도에 대한 허용 범위를 정의한다. 위의 최적화 문제는 제어 변수들에 대해서 볼록(convex)하지 않으며, 비선형성등으로 인해 일반적인 방법으로는 닫힌 형태의 해를 도출할 수 없다.

3. 제안하는 최적 UAV 제어 기법

3.1 MDP Formulation

원래 문제 (20)-(25)를 아래와 같이 MDP 문제로 재구성한다.

상태 공간(S): 시간(슬롯) t에서 상태는 다음과 같이 정의된다.

$$s_t = (x_0(t), y_0(t), (x_k, y_k)_{k=1}^{N}, R_c(t-1))$$

여기서, (x_0, y_0)는 UAV의 좌표, (x_k, y_k)는 지상 사용자의 고정된 위치, $R_c(t-1)$는 이전 슬롯에서의 공통 메시지 전송률 의미한다.

동작 공간(A): 시간 t에서의 동작은 다음과 같이 정의된다.

$$a_t = (P(t), c(t), \psi_{0(t)}, V_o(t))$$

여기서, $P(t)$는 빔포밍 행렬이며, $c(t)$는 공통 메시지 비율, $\psi_0(t)$는 UAV의 이동 방향각, $V_0(t)$는 UAV의 이동 속도를 의미한다.

보상 함수(R): 보상 함수 $R: S \times A$는 원래 문제의 목표 함수를 기반으로 정의되며, 여기서는 각 시간 t에서 다음의 보상을 제공한다.

$$r(t) = \min_{n\in N} R_{n,t} \quad (26)$$

여기서 $\min_{n\in N} R_{n,t}$는 시간 t에서 모든 사용자들 사이에서 최소 전송률을 나타낸다.

3.2 PPO 기반 UAV 전송 제어 알고리즘

본 논문에서는, 위 MDP 문제를 이용하여 최적 해를 찾는 PPO 기반 딥강화학습 (Deep Reinforcement Learning, DRL) 기법을 제안한다. 제안하는 PPO의 액터(actor)-비판(critic) 프레임워크에서, 액터 네트워크는 상태 $s(t)$에 기반하여 행동 $a(t)$를 선택하는 정책 $\pi_\theta(a(t)|s(t))$를 정의한다. 본 논문에서, UAV의 정책 공간은 연속 공간을 가정하여, 가우시안 분포 $(\mu(t), \tau^2(t))$로 설계되었다. 이 정책 함수는 θ로 매개변수화 되며, 학습 가능한 매개변수를 조정하여 주어진 상태에 대한 아래의 최적 행동을 선택할 수 있도록 한다. 첫째, n번째 빔포밍 행동 p_n는 다음과 같이 정의된다.

$$\boldsymbol{p_n} = [\alpha_1, \alpha_2, \ldots, \alpha_{2L-1}, \alpha_{2L}]^T \in R^{2L} \quad (27)$$

여기서 각 α_i는 [0, 1] 사이의 값을 갖는다. p_n은 각 스트림 l 에 대해서, 아래의 실수부와 허수부로 구성된, 스케일링된 값을 갖는다.

$$\Re(p_{n,l}) = \frac{P_t}{\sqrt{\Upsilon_n}}\alpha_{2l-1}, \Im(p_{n,1}) = \frac{P_t}{\sqrt{\Upsilon_n}}\alpha_{2l} \quad (28)$$

$$\Upsilon_n = \sum_{l=1}^{2L} \alpha l, \forall n \in N \quad (29)$$

여기서, $p_{n,l}$은 빔포밍 벡터 p_n에서 α_{2l-1}의 실수부와 α_{2l}의 허수부로 구성되는 l번째 메시지 스트림을 위한 빔포밍 벡터를 의미한다. 둘째, 공통 메시지 전송률 행동 c에 대해, N개 요소의 합이 1이 되도록 소프트맥스 함수가 적용된다.

$$\bar{c} = [\beta_1, \ldots, \beta_N]^T \in R^N \quad (30)$$

공통 메시지 전송률의 실제 스케일링된 값은 다음과 같이 계산된다.

$$R_c = \min(R_1^c, \ldots, R_N^c)\left(\sum_{n=1}^{N}\beta_n\right) \quad (31)$$

셋째, UAV의 속도 및 방향 각도 함수는 다음과 같이 계산된다.

$$V_0 = \overline{V_0}V_{max}, \varphi_0 = \bar{\varphi}_0 2\pi \quad (32)$$

종합하면, 각 단계에서 정책에 의해 결정된 동작은 다음과 같이 재정의된다.

$$\bar{a}(t) = \{\alpha_1(t), ..., \alpha_{2L}(t), \beta_1(t), ..., \beta_N(t), \overline{V_0}(t), \overline{\varphi_0}(t)\} \quad (33)$$

액터 네트워크의 정책 매개변수를 업데이트하기 위해 과거 경험 $B = \{s(t), \alpha(t), r(t), s(t+1)\}$의 배치로부터 샘플링하여 이용한다. 여기서 $s(t+1)$은 상태 $s(t)$에서 작업 $a(t)$를 수행한 후의 다음 상태를 나타낸다.

한편, 크리틱 네트워크도 액터 네트워크와 동일한 입력을 받고 상태 값을 추정한다. 크리틱 네트워크는 액터 네트워크와 독립적으로, 매개변수 ϕ를 통해서 훈련된다. 제안하는 시스템에서, 크리틱 네트워크와 액터 네트워크는 협업하여, 다음의 목적식 값을 최대화 하는 것을 목표로 한다. (i) 정책 업데이트를 안정화하기 위한 클리핑된 대리 목적 함수 (surrogate objective), (ii) 정확한 상태 가치 추정을 위한 가치 함수 손실, (iii) 그리고 탐색을 장려하기 위한 엔트로피 보너스이다. 이들이 결합된 목적식 함수는 다음과 같이 표현되며, 이를 최대화하는 것을 목표로 한다.

$$L_t^{CLIP+VF+S}(\theta) = \hat{E}[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\phi) + c_2 S[\pi_\theta(s(t))]] \quad (34)$$

여기서, 첫째로, $L_t^{CLIP}(\theta)$은 클리핑된 대리 목표로서, 아래와 같다.

$$L_t^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (35)$$

식 (35)에서, 비율 $r_t(\theta) = \dfrac{\pi_\theta(a(t)|s(t))}{\pi_{\theta_{old}}(a(t)|s(t))}$는 새로운 정책과 이전 정책을 비율을 의미하며, 클리핑된 목적 함수는 $r_t(\theta)$를 $[1-\epsilon, 1+\epsilon]$범위로 제한한다. 이 제한은 지나치게 큰 업데이트를 방지하여 정책 업데이트의 안정성을 향상시킨다 [23]. 그리고, (35)에서, 우리는 [24]에서 제안된 일반화 어드밴티지 추정(GAE) 방법을 채택하였다.

$$\hat{A}_t = \delta_t + (\gamma\iota)\delta_{t+1} + \cdots + (\gamma\iota)^{T-t-1}\delta_{T-1} \quad (36)$$

여기서 δ_t는 시간차(TD) 오차, γ는 할인 계수, ι는 어드밴티지 예측에서 bias-variance를 조절하는 GAE 가중치 변수를 나타낸다. T-단계 보상은 순환 네트워크 구조의 대안으로 사용되며, 이는 에이전트가 여러 단계에 걸쳐 미래 보상을 고려하고 누적 보상을 극대화하는 전략을 학습할 수 있도록 한다. 둘째, L_t^{VF}는 $(V_\phi(s(t)) - V_t^{target})^2$로 정의되는 제곱 오차 손실을 나타낸다. 여기서, $V_\phi(s(t))$ 와 V_t^{target}은 어드밴티지 함수 $\widehat{A_t}$의 추정되는 상태 값과 실제 보상값을 각각 나타낸다. 셋째, 엔트로피 보너스 S는 탐색을 강화하는 역할을 한다. 여기서, c_1과 c_2는 오차 손실 최소화와 탐색 강화의 균형을 맞추는 가중 계수이다.

위 전체 목적함수 (34)에 대해서, 액터 네트워크는 다음과 같이 클리핑된 손실 L^{CLIP}을 사용하여 매개변수 θ를 업데이트한다.

$$\theta_{new} = \arg\max_\theta \mathbb{E}_t\left[L_t^{CLIP}(\theta) - c_1 \cdot S[\pi_\theta(s(t))]\right] \quad (37)$$

이 손실 함수를 최대화함으로써 액터 네트워크의 매개변수 θ가 업데이트된다. 반면에, 크리틱 네트워크의 매개변수 ϕ는 L^{VF}를 최소화함으로써 업데이트된다.

$$\phi_{new} = \arg\min_\phi \mathbb{E}_t\left[\left(V_\phi(s(t)) - V_t^{target}\right)^2\right] \quad (38)$$

액터 및 크리틱 네트워크는 두 개의 은닉층을 가진 다층 퍼셉트론(MLP) 구조를 공유한다. 각 층에는 학습 안정성을 향상시키기 위해 층정규화 (layer normalization)와 ReLU 활성화가 적용된다. 크리틱 네트워크는 GAE 기반 가치 추정을 위한 단일 선형 출력으로 끝난다. 반면, 액터 네트워크는 출력 범위를 제한하기 위해 시그모이드

Algorithm 1 Proximal Policy Optimization (PPO) with Actor-Critic Structure in RSMA Environment

Initialize $\gamma, \iota, lr, \epsilon, B, D, \theta, \phi$
for episode = 1, ..., E **do**
  Observe initial state $s[0]$
  for step = 1...S **do**
    Observe state $s[t]$
    Select overall action $\bar{a}(t)$
    Execute $\bar{a}(t)$ and observe $r(t)$ and $s(t+1)$
    Store $(s(t), a(t), r(t), s(t+1))$ in replay buffer
  end for
  Calculate $\hat{A}_t$ using GAE
  Compute $L^{CLIP}(\theta)$ and $L^{VF}(\phi)$
  Update policy and update θ and ϕ using optimizer
end for
Return optimized policy parameters θ^*)

활성화로 종료되며, 액션 분산 모델링을 위해 학습 가능한 로그-표준편차를 유지한다. 제안된 PPO 알고리즘의 전체 흐름은 알고리즘 1에 제시되어 있다.

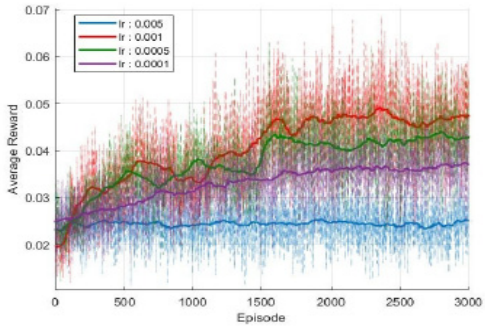
4. 성능 평가

본 장에서는 제안된 알고리즘의 성능을 아래의 알고리즘들과 비교 평가하였으며, 시뮬레이션 파라미터는 이전 연구 [22, 23]를 참조하여 결정되었다.

- DDPG: 이 기법은 제안하는 알고리즘이 동작하는 환경과 동일한 환경에서, 제안하는 알고리즘 대신 DDPG 알고리즘을 적용한 것으로, 각 에이전트는 3000 에피소드에 걸쳐, UAV 이동 경로, 빔포밍 벡터, 공통 메시지 전송률 할당을 최적화하도록 훈련되었다.
- SAC: 이 기법은 제안하는 알고리즘이 동작하는 환경과 동일한 환경에서 SAC 알고리즘을 적용하였으며, 각 에이전트는 3000회 에피소드 동안 학습되었다.
- TRPO: 이 기법은 제안하는 알고리즘이 동작하는 환경과 동일한 환경에서 TRPO 알고리즘을 적용하였으며, 각 에이전트는 3000회 에피소드 동안 학습되었다.
- REINFORCE: 이 기법은 제안하는 알고리즘이 동작하는 환경과 동일한 환경에서 REINFORCE 알고리즘을 적용하였으며, 각 에이전트는 3000회 에피소드 동안 학습되었다.
- Greedy: 이 기법은 제안하는 알고리즘이 동작하는 환경과 동일한 환경에서 UAV 이동 경로, 빔포밍 벡터, 공통 메시지 전송률 할당을 순서대로 탐욕적인 액션을 취하는 알고리즘을 적용하였으며, 각 에이전트는 3000회 에피소드 동안 학습되었다.
- Random: 이 기법은 각 시간 단계마다 UAV 이동 경로, 빔포밍 벡터, 공통 메시지 전송률을 무작위로 선택하였다.

4.1 수렴분석

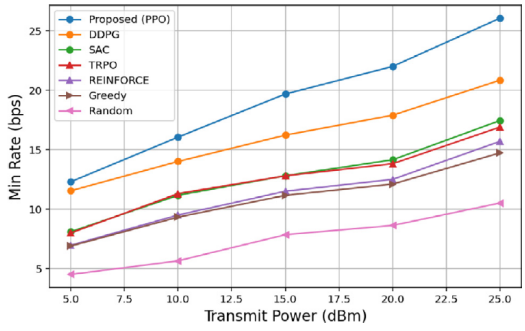
그림 2는 다양한 학습률 0.0001, 0.0005, 0.001, 0.005에 대한 학습 에피소드별 평균 보상값이 어떻게 수렴하는지를 보여준다. 위 실험에서, 검증된 값들 중 학습률 0.001이 가장 안정적이고 높은 보상 수렴을 달성함을 보였고, 학습률 0.005는 학습 행동이 불안정하게 나타나고, 평균 보상이 가장 낮은 값으로 수렴함을 보여주었다. 따라서, 본 실험에서는 이를 토대로, 학습률 0.001을 이용하여, 성능을 평가하였다.



[Fig. 2] Convergence

4.2 DRL 알고리즘 비교

그림 3은 전송 전력이 5 dBm에서 25 dBm으로 증가함에 따라 달성할 수 있는 최소 사용자 전송률을 보여준다. 제안하는 PPO 기법은 모든 전력 수준에서 DDPG, SAC, TRPO, REINFORCE, Greedy, Random 기법들보다 우수한 성능을 보여주었다. 25dBm에서 PPO는 약 26 bps의 최소 전송률을 달성한 반면에, DDPG, SAC, TRPO, REINFORCE, Greedy 기법들은 대략 21, 17, 17, 16, 15 bps의 전송률을 획득하였다. Random 정책은 랜덤 정책은 훨씬 더 낮은 성능을 보이며, 더 높은 전력 수준에서도 11 bps 미만을 유지하였다. 이러한 결과는, 제안하는 PPO 정책이 UAV 기반 RSMA 시스템과 같이 매우 복잡한 네트워크 환경에서의 복잡한 제어에도 매우 효과적으로 동작함을 보여준다.



[Fig. 3] Minimum Rate vs. Power

5. 결론

본 논문에서는 URA 안테나가 장착된 UAV- RSMA 하향링크 시스템에서, UAV 이동 경로, RSMA 빔포밍 행렬, 및 공통 메시지 전송률 최적화를 통해, 사용자 공정성(fairness)을 최대화하는 PPO 기반 제어 알고리즘을 제안하였다. 제안하는 PPO 알고리즘은, UAV가 정확한 채널 상태 정보(CSI)에 의존하지 않고도, 동적으로 변하는 환경에서도 효율적으로 간섭 관리 및 자원 할당을 가능하도록 하였다. 시뮬레이션을 통해, 제안하는 PPO 기반 제어 방식이, DDPG, SAC, TRPO, REINFORCE, Greedy, Random 과 같은 주요 방식들 보다 훨씬 향상된 사용자 최소 전송률을 제공함을 확인하였다. 향후 연구에서는, 제안하는 알고리즘을 다중 UAV 협력 및 이동 사용자 환경에서 적용 가능하도록 확장할 것이다.

REFERENCES

[1] Z. Yao, W. Cheng, W. Zhang and H. Zhang, "Resource allocation for 5g-UAV-based emergency wireless communications," IEEE Journal on Selected Areas in Communications, Vol. 39. No. 11, pp. 3395–3410, 2021.

[2] C. Liu, M. Ding, C. Ma, Q. Li, Z. Lin, Y.-C and Liang, "Performance analysis for practical unmanned aerial vehicle networks with LoS/NLoS transmissions," in 2018 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6, 2018.

[3] A. A. Nasir, H. D. Tuan, T. Q. Duong and H. V. Poor, "UAV-enabled communication using NOMA," IEEE Transactions on Communications, Vol. 67, No. 7, pp. 5126–5138, 2019.

[4] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. LI and, K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), pp. 1–5. 2013..

[5] Y. Mao, B. Clerckx and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA" EURASIP Journal on Wireless Communications and Networking, 133, 2018

[6] B. Clerckx, Y. Mao, R. Schober and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel," IEEE Wireless Communications Letters, vol. 8, no. 3, pp. 9-11, 2019.

[7] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, G. Caire, "Is NOMA efficient in multi-antenna networks? a critical look at next generation multiple access techniques," IEEE Open Journal of the Communications Society, vol. 2, pp. 1310–1343. 2021.

[8] A. Rahmati, Y. Yapici, N. Rupasinghe, I. Guvenc, H. Dai and A. Bhuyan, "Energy efficiency of RSMA and NOMA in cellular-connected mmwave UAV networks," in IEEE International Conference on Communications Workshops (ICCWorkshops), pp. 1–6, 2019.

[9] X. Liu, J. Feng, F. Li and V. C. Leung, "Downlink energy efficiency maximization for RSMA-UAV assisted communications," IEEE Wireless Communications Letters, Vol. 13, No. 1, pp. 98–102, 2023

[10] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios and H. Yanikomeroglu, "On the downlink performance of RSMA-based UAV communications," IEEE Transactions on Vehicular Technology, Vol. 69, No. 12, pp. 16258–16263, 2020

[11] B. Yao, R. Li, Y. Chen and L. Wang, "Coordinated RSMA for integrated sensing and communication in emergency UAV systems," arXiv:2406.19205, 2024.

[12] O. Dizdar, A. Sattarzadeh, Y. X. Yap and S.Wang, "RSMA for overloaded MIMO networks: Low-complexity design for max–min fairness," IEEE Transactions on Wireless Communications, Vol. 23, No. 6, pp. 6156–6173, 2024.

[13] B. Lee and W. Shin, "Max-min fairness precoder design for rate-splitting multiple access: Impact of imperfect channel knowledge," IEEE Transactions on Vehicular Technology, Vol 72, No. 1, pp. 1355–1359, 2023.

[14] Y. Xu, Y. Mao, O. Dizdar and B. Clerckx, "Max-min fairness of rate-splitting multiple access with finite blocklength communications," IEEE Transactions on Vehicular Technology, Vol. 72, No. 5, pp. 6816–6821, 2021.

[15] M. Xiao, H. Cui, D. Huang, Z. Zhao, X. Cao and D. O.Wu, "Traffic-aware energy-efficient resource allocation for RSMA based UAV communications," IEEE Transactions on Network Science and Engineering, Vol. 11, No. 3, pp. 2537–2548, 2023.

[16] J. Feng, X. Liu, Z. Liu and T. S. Durrani, "Optimal trajectory and resource allocation for RSMA-UAV assisted IoT communications," IEEE Transactions on Vehicular Technology, Vol. 74, No. 10, pp. 8693–8704, 2024.

[17] L. Huang, S. Bi, Y.-J and A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," IEEE Transactions on Mobile Computing, Vol. 19, No. 11, pp. 2581–2593. 2020.

[18] Y. Sun, M. Peng and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," IEEE Internet of Things Journal, Vol. 6, No. 2, pp.

1960-1971, 2019.

[19] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun and H. Gao, "Cooperative UAV trajectory design for disaster area emergency communications: A multiagent PPO method," IEEE Internet of Things Journal, Vol. 11, No. 5, pp. 8848-8859, 2023.

[20] G. Zhan, X. Zhang, Z. Li, L. Xu, D. Zhou, Z and Yang, "Multiple-UAV reinforcement learning algorithm based on improved ppo in ray framework," Drones 6, No. 7: 166, 2022.

[21] S. K. Yong and J. S. Thompson, "Three-dimensional spatial fading correlation models for compact MIMO receivers," IEEE Transactions on Wireless Communications, Vol. 4, No. 6, pp. 2856-2869, 2005.

[22] C. You and R. Zhang, "3D trajectory optimization in rician fading for UAV-enabled data harvesting," IEEE Transactions on Wireless Communications, Vol. 18, No. 6, pp. 3192-3207, 2019.

[23] M. Hata, "Empirical formula for propagation loss in land mobile radio services," IEEE transactions on Vehicular Technology, Vol. 29, No. 3, pp. 317-325, 1980.

노 원 종(Wonjong Noh) [정회원]

■ 2024년 3월 ~ : 한림대학교 소프트웨어학부 교수
■ 2007년 3월 ~ : 삼성전자 종합기술원 수석연구원
■ 2005년 3월 ~: Purdue Univ. Post. Doc.

■ 2005년 2월 : 고려대학교 전자공학과 박사
■ 2000년 2월 : 고려대학교 전자공학과 석사
■ 1998년 2월 : 고려대학교 전자공학과 학사

〈관심분야〉
이동통신 시스템, 최적 제어, 분산 컴퓨팅, 시맨틱 전송 인공지능 및 머신러닝