불완전한 CSI 환경에서의 NOMA 자원 할당을 위한 심층 강화학습 기법

문찬호, 오민택, 최진석
한국과학기술원

{ghcksans, ohmin, jinseok}@kaist.ac.kr

Deep RL for NOMA Resource Allocation under Imperfect CSI

Chanho Moon, Mintaek Oh, Jinseok Choi
Korea Advanced Institute of Science and Technology

요 약

본 논문은 채널의 동특성(Dynamics)을 사전에 알 수 없고(Unknown MDP), 채널 상태 정보(CSI)마저 불완전한(POMDP) 하향링크 비직교 다중 접속(NOMA) 시스템을 다룬다. 이러한 복합적인 불확실성 환경에서 사용자 공정성과 스펙트럼 효율을 동시에 최적화하기 위해, 에이전트가 채널의 확률 분포를 추정하여 불확실성을 정량화하는 '확률 상태 인지형' 심층 강화학습(DRL) 프레임워크를 제안한다. 제안 기법은 미지의 채널 변화 패턴을 학습함과 동시에, 불확실한 관측 정보를 확률적으로 보정하여 의사결정에 활용함으로써 기존 기법 대비 강인한 성능을 입증하였다.

Ⅰ. 서 론

6G 및 IoT 네트워크를 위한 비직교 다중 접속(NOMA) 기술은 높은 주파수 효율을 제공하지만, 최적 자원 할당은 NP-hard 문제에 속한다 [1]. 실제 무선 환경에서의 자원 할당은 두 가지 핵심적인 난제에 직면한다. 첫째, 페이딩 채널의 복잡한 천이 확률(Transition Dynamics)을 사전에 알 수 없는 'Unknown MDP' 문제이다 [2]. 둘째, 피드백 지연과 추정 오차로 인해 기지국이 현재의 정확한 상태를 관측할 수 없는 '부분 관측 마르코프 결정 과정(POMDP)' 문제이다 [3]. 기존 최적화 기법이나 단순한 강화학습은 완벽한 CSI 를 가정하거나 환경의 동특성을 충분히 고려하지 못하여 성능이 저하된다. 이에 본 논문에서는 미지의 채널 동특성을 데이터로부터 학습하고, 동시에 불완전한 관측 정보를 확률적 믿음(Belief) 상태로 변환하여 POMDP 문제를 해결하는 새로운 DRL 프레임워크를 제안한다.

Ⅱ. 본론

본 논문에서는 단일 기지국(BS)이 N 명의 사용자와 K 개의 직교 부채널을 통해 통신하는 하향링크 NOMA 시스템을 고려한다. 시간 t 에서의 송신 신호 $x_k(t)$ 는 중첩 코딩(Superposition Coding)되어 전송되며, 사용자 n 이 부채널 k 에서 수신하는 신호 $y_{n,k}(t)$ 는 채널 이득 $h_{n,k}(t)$와 가산 백색 가우시안 잡음 $z_{n,k}(t)$에 의해 다음과 같이 결정된다 [4].

$$y_{n,k}(t) = h_{n,k}(t) \sum_{j \in \mathcal{N}_k(t)} \sqrt{p_{j,k}} s_{j,k}(t) + z_{n,k}(t).$$

수신단에서는 순차적 간섭 제거(SIC)를 수행하며, 이때 사용자 n의 신호 대 간섭 잡음비(SINR)는 다음과 같다.

$$\mathrm{SINR}_{n,k}(t) = \frac{p_{n,k}(t)\left|h_{n,k}(t)\right|^2}{\sum_{j \in \mathcal{N}_k(t), |h_j|^2 > |h_n|^2} p_{j,k}(t)\left|h_{n,k}(t)\right|^2 + \sigma^2}.$$

본 연구의 목표는 채널 할당 지시자 $u_{n,k}(t) \in \{0,1\}$와 전력 할당 $p_{n,k}(t)$ 를 최적화하여 시스템 총 전송률과 공정성을 최대화하는 것이며, 이는 다음과 같은 최적화 문제로 정식화된다 [7].

$$\max_{\{u,p\}} \sum_{n=1}^{N} R_n(t) - \zeta\left(1 - J(t)\right)$$

$$\text{subject to} \sum_{k=1}^{K}\sum_{n=1}^{N} u_{n,k}(t) p_{n,k}(t) \le P_{max},$$

$$p_{n,k}(t) \ge 0,$$

$$\sum_{k=1}^{K} u_{n,k}(t) = 1.$$

여기서$J(t)$는 Jain's Fairness Index 로, 각 사용자의 시간 평균 전송률 $\bar{R}_n(t)$ 를 기반으로 $\mathrm{J(t)} = \frac{(\sum \bar{R}_n(t))^2}{N \sum (\bar{R}_n(t))^2}$ 와 같이 계산된다 [7]. 이때 평균 전송률은 $\bar{R}_n(t) = (1-\alpha)\bar{R}_n(t-1) + \alpha R_n(t)$ 로 업데이트되므로, 현재의 공정성 계산에는 과거의 전송 이력이 필수적이다. 따라서 본 문제는 순차적 의사결정 문제(Sequential Decision Process)가 되며, 기지국이 환경의 동특성을 모르는(Unknown MDP) 동시에 피드백 지연으로 인해 정보가 불완전한 상황을 반영하여 부분 관측 마르코프 결정 과정(POMDP)으로 정의된다 [3].

구체적으로 본 POMDP 는 튜플 $\langle S, O, A, R \rangle$로 정의된다. 먼저 상태 $S_t = \{g_t, \bar{R}(t), A_{t-1}\}$ 는 기지국이 관측할 수 없는 실제 채널 이득 벡터 g_t 와 마르코프 성질을 만족시키기 위한 평균 전송률 $\bar{R}(t)$, 이전 행동 A_{t-1}을 포함한다. 반면, 에이전트가 획득하는 관측 O_t 는 과거 L 개의 관측된 채널 이득 목록 $\hat{g}_t$ 와 해당 정보의 경과 시간 age(t) 등으로 구성된다. 즉, $\mathrm{O_t} = \{\hat{g}_t, \mathrm{age(t)}, \bar{R}(t), \mathrm{A_{t-1}}\}$ 이며, 여기서 $\hat{g}_t$ 는 단순히 현재 시점의 값이 아닌 과거 관측치들의 시퀀스를 포함하여 에이전트가 채널의 시계열적 패턴을 추론할 수 있도록 한다. 이에 대응하는 행동 $A_t = \{u(t), p(t)\}$는 이산적인 채널 할당과 연속적인 전력 할당을 동시에 결정하는 결합 벡터이며, 보상 R_t 는 최적화 목표와 동일하게 $\sum_{n=1}^{N} R_n(t) - \zeta\left(1 - J(t)\right)$로 정의되어 에이전트가 효율성과 공정성 간의 트레이드오프를 학습하도록 유도한다[7].

이러한 POMDP 문제를 해결하기 위해, 본 논문은 PPO(Proximal Policy Optimization) [6] 기반의 에이전트에 LSTM 을 결합하고, 확률 분포 예측 보조 작업(Auxiliary Task)을 도입한다. 에이전트의 LSTM

네트워크는 과거의 관측 이력 H_t를 입력받아 시계열적 특성을 학습하며, 다음 시점의 채널 이득 $|h_{n,k}(t+1)|^2$이 따르는 확률 분포의 파라미터 θ_{pred}(예: Rician 분포의 ν, σ)를 예측한다 [5]. 이 예측을 학습하기 위한 보조 손실 함수 L_{NLL}은 실제 채널 값에 대한 Negative Log-Likelihood로 정의된다.

$$L_{NLL} = -\sum_{n,k} \log P\left(|h_{n,k}(t+1)|^2 \middle| \theta_{pred}\right).$$

최종적으로 에이전트를 학습시키기 위한 전체 손실 함수 L_{Total}은 PPO 알고리즘의 손실 함수 L_{PPO}와 보조 작업의 손실 함수 L_{NLL}의 가중 합으로 구성된다.

$$L_{Total} = L_{PPO} + \beta L_{NLL}.$$

이를 통해 에이전트는 단순한 채널 값 예측을 넘어, 예측의 불확실성을 인지하고 이를 정책에 반영하여(Risk-aware) 강인한 자원 할당을 수행하게 된다.

성능 검증을 위해 3GPP UMi 채널 모델을 기반으로 $N = 6$, $K = 3$, $P_{max} = 30$ dBm 환경에서 모의실험을 수행하였다. 제안 기법의 우수성을 검증하기 위해 다음 세 가지 비교군(Baseline)을 설정하였다. 첫째, SR-1 과 SR-2 는 각각 가중 합 전송률 최대화와 최소 QoS 보장을 목표로 하는 전통적인 최적화 휴리스틱이다. 이들은 미래 채널을 예측하지 못하고 현재의 관측 정보에만 의존하며, 채널 당 사용자 수가 최대 2 명으로 제한된다는 한계가 있다 [8]. 둘째, RL-LSTM(MSE)은 제안 기법과 동일한 신경망 구조를 가지나, 보조 작업으로 확률 분포가 아닌 다음 시점의 채널 값 자체를 예측(점 추정)하며 평균 제곱 오차(MSE) 손실 함수를 사용하는 모델이다.

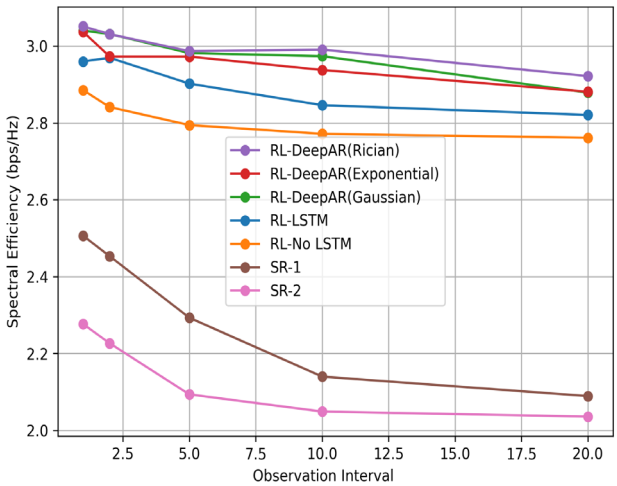

그림 1. 관측 주기에 따른 스펙트럼 효율

[그림 1]과 [그림 2]는 CSI 관측 주기 K_{obs}가 증가함에 따른 스펙트럼 효율과 공정성 변화를 보여준다. 환경의 동특성을 학습하지 못하는 기존 휴리스틱(SR-1, SR-2) 및 Memoryless RL 은 관측 주기가 길어짐에 따라 성능이 급격히 저하된다. 또한, 점 추정을 수행하는 RL-LSTM(MSE)은 제안 기법보다 낮은 성능을 보이는데, 이는 불확실한 환경에서는 단순한 값 예측보다 분포 추정을 통한 위험 관리(Risk Management)가 필수적임을 시사한다. 특히, 채널의 통계적 특성(LoS/NLoS 혼합)을 가장 잘 반영하는 Rician 분포 기반의 RL-DeepAR(Rician) 모델이 실제 채널 분포와의 불일치를 최소화하여, Gaussian 이나 Exponential 모델 대비 가장 뛰어난 스펙트럼 효율과 공정성을 달성함을 확인하였다.

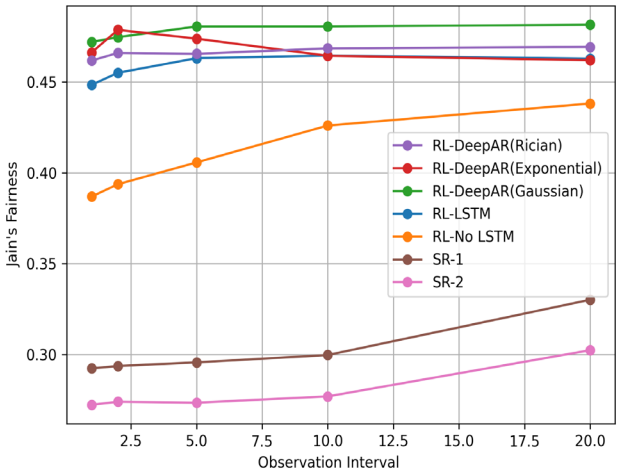

그림 2. 관측 주기에 따른 Jain's Fairness Index

Ⅲ. 결론

본 논문에서는 Unknown MDP 및 POMDP 특성을 가진 NOMA 자원 할당 문제를 해결하기 위해, 환경의 동특성을 학습하고 불확실성을 정량화하는 확률 상태 인지형 DRL 프레임워크를 제안하였다 . 모의실험 결과, 제안 기법은 불완전한 정보 하에서도 우수한 성능과 강인함을 입증하였으며, 이는 차세대 통신 시스템의 자원 관리에 효과적으로 적용될 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 지원(No. RS-2025-20552984, 50%)과 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2024-00395824, (총괄 1-세부 2) Upper-mid Band를 지원하는 Cloud virtualized RAN (vRAN) 시스템 기술 개발, 50%)

참 고 문 헌

[1] L. Salaün et al., "Optimal Joint Subcarrier and Power Allocation in NOMA is Strongly NP-Hard," ICC, 2018.

[2] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications," IEEE Comm. Surveys & Tuts., vol. 21, 2019.

[3] L. P. Kaelbling et al., "Planning and acting in partially observable stochastic domains," Artificial Intelligence, vol. 101, 1998.

[4] Z. Ding et al., "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," IEEE Comm. Mag., vol. 55, 2017.

[5] D. Salinas et al., "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks," Int. J. Forecasting, vol. 36, 2020.

[6] J. Schulman et al., "Proximal Policy Optimization Algorithms," arXiv:1707.06347, 2017.

[7] R. Jain et al., "A quantitative measure of fairness and discrimination," DEC Research Report TR-301, 1984.

[8] J. Zhu et al., "On Optimal Power Allocation for Downlink Non-Orthogonal Multiple Access Systems," IEEE JSAC, vol. 35, no. 12, 2017.