무인 항공기 기반 통신에서 사용자 스케줄링을 위한 제약 조건 기반 강화학습

김정원 [1], 황상원 [2], 김상민 [1], 부지림 [3], 문지환 [4], 이인규 [1*]
[1] 고려대학교, [2] 국립부경대학교, [3] 삼성전자, [4] 국립한밭대학교

{jeongwonkim, smgeem, inkyu}@korea.ac.kr, s.won.hwang@pknu.ac.kr,
zhilin.fu@samsung.com, anschino@staff.hanbat.ac.kr

Constrained Deep Reinforcement Learning for User Scheduling in UAV-Enabled Communications

Jeongwon Kim[1], Sangwon Hwang[2], Sangmin Kim[1], Zhilin Fu[3], Jiwhan Moon[4], Inkyu Lee[1*]

[1]Korea Univ., [2]Pukyong National Univ., [3] Samsung Electronics Co., [4]Hanbat National Univ.

요 약

본 논문은 UAV 기반 통신에서 통신 효율을 높이면서 사용자별 QoS 제약을 안정적으로 만족시키기 위한 constrained RL 프레임워크를 제안한다. 기존 강화학습은 QoS 와 같은 제약조건을 학습 과정에서 안정적으로 보장하기 어렵고, penalty 설계에 민감해 성능이 불안정해지는 한계가 있다. 제안한 프레임워크는 목적함수와 제약을 분리해 다루는 최적화 관점을 강화학습에 결합함으로써 이러한 문제를 완화하고, 기존 DDPG 대비 성능 이득을 확인하였다.

Ⅰ. 서론

강화학습을 실제 시스템에 적용하는 과정에서 제약 조건을 안정적으로 만족해야 하는 문제가 필연적으로 발생한다. penalty 기반 접근은 reward engineering 에 대한 민감도가 매우 높고, penalty 가중치에 따라 학습이 불안정해지거나 지나치게 보수적인 정책으로 수렴하는 문제가 발생할 수 있다. 또한, 환경 변화에 따라 제약 만족이 쉽게 깨질 수 있고, 또한 제약 위반이 드물게 발생하거나 특정 상태에서만 나타나는 경우에는 agent 가 제약을 만족하는 정책을 학습하기까지 많은 episode 가 필요하다. 이러한 한계를 보완하기 위해 성능 최적화와 제약 만족을 동시에 다루는 CMDP, safe Reinforcement Learning(RL), constrained RL 과 같은 프레임워크가 제안되어 왔다. [1]

본 논문에서는 무인 항공기 기반 통신에서 통신 효율을 높이면서 각 사용자에 대한 QoS 를 만족시키기 위해 목적함수와 제약을 분리하여 다루는 최적화 관점을 강화학습에 결합하는 프레임워크를 제안한다.

Ⅱ. 본론

본 논문에서는 UAV가 지상에 분포한 K명의 사용자를 대상으로 통신 서비스를 제공하는 UAV 기반 통신 시나리오를 고려한다. UAV 가 이동하는 전체 시간 구간은 N 개의 timeslot 으로 분할하며, 각 timeslot 의 시간은 δ_t로 일정하다. UAV 는 고도 H에서 비행한다고 가정하며, timeslot n에서의 UAV 위치는 $\boldsymbol{q}[n] = (q_x[n], q_y[n], H)$로 가정한다. 또한 k번째 사용자의 위치는 $\boldsymbol{u}_k = (u_{k,x}, u_{k,y}, 0)$로 정의한다. 각 timeslot n에서 UAV 가 사용자 k 에게 서비스를 제공하는지 여부는 binary variable $c_k[n] \in \{0,1\}$ 로 정의하며, TDMA 기반 스케줄링을 가정하여 timeslot n 에서 동시에 서비스 가능한 사용자는 최대 1 명으로 제한한다. 이에 따라 timeslot n에서의 사용자 선택 제약은 $\sum_{k=1}^{K} c_k[n] \leq 1$로 나타낸다.

실제 시나리오에서 주변 구조물 정보가 부족하기 때문에 채널 모델에 line-of-sight(LoS) 및 non-line-of-sight(NLoS) 링크의 무작위성을 고려한다. [3,4] 이때 timeslot n에서 UAV 와 k번째 user 사이의 LoS 확률은 다음과 같이 나타낸다.

$$P_{LoS}^{k}[n] = \frac{1}{1 + K_1 \exp\left(-K_2(\theta_k[n] - K_1)\right)}$$

K_1과 K_2는 반송파 주파수 및 농촌, 도심, 밀집 도심 등과 같은 전파 환경 유형에 따라 결정되는 상수이며, θ_k 는 UAV 와 사용자 k 사이의 고도각이다.

UAV 는 single omnidirectional antenna 를 장착하였다고 가정하고, TDMA 를 가정하므로 inter-beam interference 는 존재하지 않는다. [5] timeslot n에서 UAV 와 사용자 k 사이의 channel gain 은 다음과 같이 나타낼 수 있다.

$$g_k[n] = \begin{cases} \dfrac{\beta_0}{\chi_{LoS}(H^2 + \|\boldsymbol{q}[n] - \boldsymbol{u}_k\|^2)}, with\ probability\ P_{LoS}^{k}[n] \\ \dfrac{\beta_0}{\chi_{NLoS}(H^2 + \|\boldsymbol{q}[n] - \boldsymbol{u}_k\|^2)}, otherwise \end{cases}$$

β_0는 기준 거리 1m 에서의 채널 전력 이득이고, χ_{LoS}와 χ_{NLoS}는 각각 LoS 및 NLoS 환경에서의 추가 경로 손실 계수이며, 이때 $\chi_{NLoS} > \chi_{LoS} > 1$이다.

UAV 와 각 user 의 위치만 주어질 때, 링크의 LoS 또는 NLoS 상태를 정확하게 알 수 없기 때문에 각 UAV 와 사용자 간의 LoS 와 NLoS 에 대한 평균 channel gain 을 고려한다.

$$\bar{g}_k[n] = P_{LoS}^{k}[n] \cdot \frac{\beta_0}{\chi_{LoS}(H^2 + \|\boldsymbol{q}[n] - \boldsymbol{u}_k\|^2)} + (1 - P_{LoS}^{k}[n]) \cdot \frac{\beta_0}{\chi_{NLoS}(H^2 + \|\boldsymbol{q}[n] - \boldsymbol{u}_k\|^2)}$$

timeslot n 에서 통신 스펙트럼 효율은 다음과 같이 나타낼 수 있다.

$$E^{com}[n] = \sum_{k=1}^{K} c_k[n] \log_2(1 + \frac{P_{max}\bar{g}_k[n]}{\sigma^2})$$

P_{max} 는 UAV 의 최대 송신 전력이고, σ^2 은 잡음 전력이다.

마찬가지로, timeslot n 에서 사용자 k 의 통신 스펙트럼 효율도 다음과 같이 나타낼 수 있다.

$$E_k^{com}[n] = c_k[n]\log_2(1 + \frac{P_{max}\bar{g}_k[n]}{\sigma^2})$$

통신 스펙트럼 효율은 각 timeslot 에서 어떤 user 를 선택할지 결정하는 user scheduling 에 따라 최대화되며, 이때, 전체 timeslot 동안 각 er 들은 quality-of-service(QoS)를 만족하기 위해서 최소 전송률 요구사항을 만족해야 한다. 따라서, 문제를 다음과 같이 정의할 수 있다.

$$(P1): \max_{\{c[n]_{n=1}^N\}} \frac{1}{N}\sum_{n=1}^{N} E^{com}[n]$$
$$s.t. \quad \sum_{n=1}^{N} E_k^{com}[n] \geq \eta^{th}, \forall k$$

이때, $\boldsymbol{c}[n] = \{c[n]_{n=1}^N\}$ 이고, η^{th} 는 최소 전송률 요구사항이다.

본 논문에서는 각 timeslot 에서 최적의 user 를 할당함으로써 통신 스펙트럼 효율을 최대화하는 동시에, 각 user 의 QoS 요구사항 같은 제약을 모든 timeslot 에 걸쳐 만족하도록 Constrained DRL 학습 구조를 설계한다.

먼저, Constraint violation 을 다음과 같이 정의한다.

$$v \triangleq \max_k(\max(0, \eta^{th} - E_k^{com}))$$

이를 통해 문제를 다음과 같이 재정의할 수 있다.

$$(P1-1): \max_{\{c[n]_{n=1}^N\}} \frac{1}{N}\sum_{n=1}^{N} E^{com}[n]$$
$$s.t. \quad v \leq 0$$

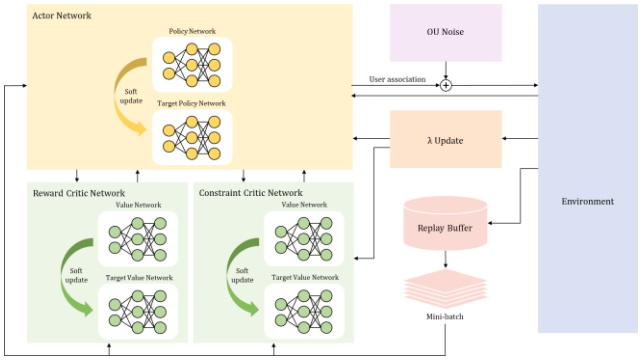

그림 1 Constrained DDPG 구조

Lagrangian relaxation 기반 방법을 사용하여 제약 위반 정도에 따라 Lagrange multiplier 를 적응적으로 업데이트하여, penalty 가중치를 수동으로 튜닝하지 않고도 제약 만족과 성능 최적화 사이의 균형을 자동으로 조절할 수 있는 framework 를 제안한다. 그림 1 은 제안하는 framework 의 구조로 constraint 를 고려하기 위해 Actor 네트워크의 loss 로 Lagrangian loss 가 도입되고, Reward 와 Constraint 에 대한 Critic Network 를 각각 설계함으로써, 새로운 DRL structure 를 제안한다. 문제에 대한 Lagrangian relaxation 은 다음과 같이 나타낼 수 있다.

$$\mathcal{L}(\boldsymbol{c}[n]_{n=1}^N, \lambda) = \frac{1}{N}\sum_{n=1}^{N} E^{com}[n] - \lambda v, \qquad \lambda \geq 0$$

Ⅲ. 결론

본 논문에서는 1 대의 UAV 가 일정한 속도로 지정된 경로를 따라 움직이며 2 명의 user 가 있는 상황을 가정하였으며, $N = 5$, $\eta^{th} = 13bits/Hz$로 설정하였다.

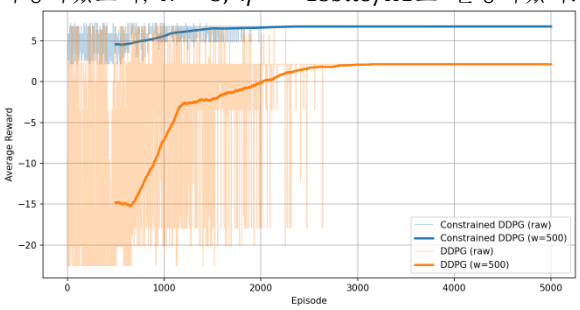

그림 2 Episode 에 따른 reward

그림 2 은 서로 다른 RL framework 에서의 episode 별 reward 를 비교한 결과이다. Constrained DDPG 는 DDPG 에 비해 더 좋은 성능을 내고 최적의 정책을 학습하기까지의 필요한 episode 가 적으며, violation 을 하지 않는 방향으로 학습함으로써 제약 조건이 존재하는 문제에서 좋은 성능을 보인다.

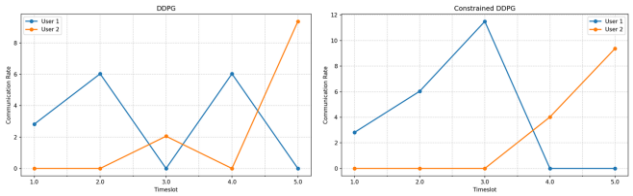

그림 3 Timeslot 에 따른 user scheduling

그림 3 에서는 동일한 환경에서 Constrained DDPG 와 DDPG 가 각 timeslot 에서 선택한 action 의 결과를 사용자별 communication rate 로 나타낸 것이다. DDPG 대비 Constrained DDPG 는 각 timeslot 에서 최적의 action 을 선택함으로써 QoS 제약을 만족하면서도 동시에 더 높은 rate 을 달성한 것을 확인할 수 있다.

ACKNOWLEDGMENT

본 연구는 한국연구재단의 지원을 받아 수행되었음. (RS-2022-NR070834).

참 고 문 헌

[1] Achiam, Joshua, et al. "Constrained policy optimization." *International conference on machine learning*. PMLR, 2017.

[2] Al-Hourani, Akram, Sithamparanathan Kandeepan, and Simon Lardner. "Optimal LAP altitude for maximum coverage." *IEEE wireless communications letters* 3.6 (2014): 569-572.

[3] Hwang, Sangwon, et al. "Deep reinforcement learning approach for UAV-assisted mobile edge computing networks." *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022.

[4] Hou, Peng, et al. "Distributed DRL-based integrated sensing, communication, and computation in cooperative UAV-enabled intelligent transportation systems." *IEEE Internet of Things Journal* 12.5 (2024): 5792-5806.