# A sum rate maximization problem in uplink MIMO with RSMA systems

Phung Truong[1*]

[1]Sejong University, Seoul, Korea
[*]Corresponding author: thanhphung2110 @gmail.com

**ARTICLE INFO**

**DOI:**10.46223/HCMCOUJS.tech.en.14.1.2955.2024

Received: September 07th, 2023

Revised: October 06th, 2023

Accepted: October 23rd, 2023

*Keywords*:

deep reinforcement learning; multiple-input multiple-output; rate splitting multiple access; sum rate maximization

**ABSTRACT**

This study explores the problem of maximizing the sum rate in uplink multi-user Multiple-Input Multiple-Output (MIMO) using Rate-Splitting Multiple Access (RSMA) systems. The investigation revolves around the scenario where the Users (UEs) are single-antenna nodes transmitting data to a multi-antenna Base Station (BS) through the RSMA technique. The optimization process encompasses determining parameters such as UEs' transmit powers, decoding order, and detection vector at the BS. An approach based on Deep Reinforcement Learning (DRL) is introduced to address this challenge. This DRL framework involves an action-refined stage and applies a Deep Deterministic Policy Gradient (DDPG)-based strategy. Simulation outcomes effectively demonstrate the convergence of the proposed DRL framework, where it converges after approximately 1,800 episodes. Also, the results prove the superior performance of the proposed method when compared to established benchmark strategies, where it is up to 45% and 86% higher than the local search and random schemes, respectively.

## 1. Introduction

The imminent scenarios in next-generation communications, including augmented reality, connected vehicles, and diverse Internet of Things (IoT) applications, demand elevated throughput capacities while adhering to stricter latency limitations than the current networks. To cater to these evolving requirements, the concept of Rate-Splitting Multiple Access (RSMA), blending the principles of non-orthogonal multiple access and space-division multiple access, has appeared as a dynamic and all-encompassing strategy for shaping the realm of multiple access control in upcoming wireless networks (Mao et al., 2022). Besides, multi-antenna systems play a crucial role in enhancing the performance of wireless communication. Notably, adopting Multiple-Input Multiple-Output (MIMO) technology, which uses multiple receiver antennas to boost data throughput significantly, holds a prominent position in wireless communication (Zheng, Wong, & Ng, 2009). This convergence of multi-antenna systems with diverse multiple access techniques creates fertile terrain for intriguing research avenues.

Numerous contemporary investigations have delved into the downlink transmission of RSMA coupled with MIMO systems. Park Choi, Lee, Shin, and Poor presented a spectral efficiency problem in an RSMA downlink MIMO system (Park, Choi, Lee, Shin, & Poor, 2023), where they considered optimizing the precoding matrix at the transmitter. Authors in de Sena et al. (2022) investigated a downlink MIMO network with a dual-polarized RSMA technique, where they modeled the polarization interference effects to overcome Successive Interference Cancellation (SIC) practical issues. Although research in RSMA MIMO systems has been discussed a lot recently, the consideration of RSMA in uplink MIMO systems remains attractive

in many aspects. Therefore, this study investigates an uplink MIMO system with the RSMA technique, which examines a spectral efficiency problem. Besides, machine learning has become a powerful tool to solve the issues of modern life (Nguyen, Park, & Park, 2023; Nguyen, Vo, & Nguyen, 2023; Tran, Bao, Nguyen, & Park, 2022). Significantly, the applications of reinforcement learning to solve communication problems are being extensively studied (Nguyen, Park, Seol, So, & Park, 2023). To keep up with that tendency, this work proposes a Deep Reinforcement Learning (DRL) framework to solve the problem in this study. In a nutshell, the main contributions of this paper are summarized as follows:

- The system investigates an RSMA MIMO system for uplink communication. This system enables multiple single-antenna Users (UEs) to share communication resources simultaneously, connecting to a multiple-antenna Base Station (BS) by applying the RSMA technique. Here, a spectral efficiency problem is formulated that maximizes the system sum rate by considering the detection vector, decoding orders, and the user's transmit powers as optimization variables.

- To solve the problem, a DRL framework that applies the Deep Deterministic Policy Gradient (DDPG) algorithm is proposed in the solution part. The simulation results show the convergence of the training process. Also, it assesses the proposed framework's performance in different environmental scenarios.

## 2. Problem statement

### *2.1. System model*

The considered system includes an M-antenna BS that serves N single-antenna UEs, where the uplink transmission from UEs to BS is conducted using the RSMA technique. Here, UE *n* splits its signal into two sub-signals and transmits it to the BS (Yang, Chen, Saad, Xu, & Shikh-Bahaei, 2022). Accordingly, the transmit signal at UE *n* is represented as (Nguyen & Park, 2023):

$$s_n = s_{n,1}\sqrt{p_{n,1}} + s_{n,2}\sqrt{p_{n,2}}, \tag{1}$$

Where $p_{n,i} \geq 0, i \in \{1,2\}$, denotes the transmit power of sub-signal $s_{n,i}$. At the BS, the composite signals, $\boldsymbol{y} \in \mathbb{C}^{M\times 1}$, is represented as:

$$\boldsymbol{y} = \sum_{n=1}^{N} \boldsymbol{h_n} s_n + \boldsymbol{n}, \tag{2}$$

Where $\boldsymbol{n} \in \mathbb{C}^{M\times 1}$ denotes the noise vector, and $\boldsymbol{h_n} \in \mathbb{C}^{M\times 1}$ is the channel gain vector between UE $n$ and the BS. The BS applies a unit norm vector $\boldsymbol{w} \in \mathbb{C}^{M\times 1}$ as the detection vector to detect the received signals (Ma, Ren, Quan, & Feng, 2022). Consequently, the sub-signal $s_{n,i}$ received at the BS is expressed as:

$$\hat{s}_{n,i} = \boldsymbol{w^H y}. \tag{3}$$

In uplink RSMA, the sub-signals received at the receiver are decoded by applying the SIC technique following a decoding order. Considering the decoding process follows ascending order, the achievable rate of sub-signal $s_{n,i}$ is calculated as:

$$r_{n,i} = C\ log_2\left(1 + \frac{p_{n,i}\left|\boldsymbol{w^H h_n}\right|^2}{\sum_{\pi_{n'i'}>\pi_{n,i}} p_{n',i'}\left|\boldsymbol{w^H h_{n'}}\right|^2 + \sigma^2}\right), \tag{4}$$

Where $\sigma^2$ denotes the noise power, $C$ is the communication bandwidth, and $\pi_{n,i}$ is the decoding order of $s_{n,i}$.

### *2.2. Problem formulation*

This study investigates the sum rate maximization problem, which optimizes the transmit power of UEs, the detection vector, and the decoding order at the BS. Then, the optimization is formulated as follows:

$$(\mathcal{P}1) \quad \max_{\{p_{n,i},\pi_{n,i}|i\in\{1,2\},n\in\{1,2,..N\}\}, \boldsymbol{w}} \quad \sum_{n=1}^{N} r_{n,1} + r_{n,2} \tag{5a}$$

$$\text{s.t.} \quad p_{n,1} + p_{n,2} \le P_n;\ p_{n,1} \ge 0, p_{n,2} \ge 0,\ n \in \{1,2,\dots,N\}, \tag{5b}$$

$$\left|\left|\boldsymbol{w}\right|\right|_2 = 1, \tag{5c}$$

Where constraint (5b) ensures the feasible range of transmit power, which cannot exceed the maximum power $P_n$ of each UE; and (5c) is the unit norm constraint of the detection matrix. To solve this problem with the combination of continuous variables and the decoding order, we convert it into a reinforcement learning-based problem and apply a DRL algorithm to train the agent to decide the variables.

## 3. Proposed solution

### *3.1. Reinforcement learning-based problem*

The problem $(\mathcal{P}1)$ is presented as an RL-based problem (Sutton & Barto, 2018) with the agent implemented at the BS, and the environment is the entire system, where the state, action, and reward at time step $t$ are defined as:

*State space*: The state space at time slot $t$ contains the channel gains between UEs and the BS, which is given as:

$$s[t] = \{\boldsymbol{h_n}[t]|\ n \in \{1,2,\dots,N\}\}. \tag{6}$$

*Action space*: The action space contains the variables the agent has to optimize, which is given as:

$$a[t] = \left\{\boldsymbol{w},\ p_{n,i},\pi_{n,i}\middle|i\in\{1,2\},n\in\{1,2,..N\}\right\}. \tag{7}$$

*Reward function*: Because as the sum rate maximization problem, the reward function is defined at the sum rate of UEs at each time slot, which is calculated as:

$$r[t] = \sum_{n=1}^{N} r_{n,1} + r_{n,2}. \tag{8}$$

Accordingly, a DRL algorithm named DDPG is applied to train the agent to decide the appropriate action at each time slot by observing the environment's state.

### *3.2. DDPG algorithm*

DDPG is an actor-critic algorithm that includes actor and critic networks (Lillicrap et al., 2015). The actor is a policy network that directly maps states to continuous actions. Its goal is to learn a policy that maximizes the expected cumulative reward. The critic is a value network that estimates the expected cumulative reward (Q-value) of being in a certain state and taking a certain action. It provides the actor with a feedback signal to guide its learning. To stabilize training, DDPG uses target networks for both the actor and the critic. These target networks are periodically updated by soft copying the weights from the primary networks. This mitigates the problem of rapidly changing Q-values during learning. The primary actor-network ($\vartheta(s|\epsilon^{\vartheta})$, $\epsilon^{\vartheta}$ is the network parameter) is trained by a gradient function, expressed as:

$$\nabla_{\epsilon^{\vartheta}} J = \frac{1}{D}\sum_{d=1}^{D} \nabla_{\vartheta(s_d)} Q\left(s_d, \vartheta\left(s_d|\epsilon^{\vartheta}\right)\middle|\epsilon^{Q}\right) \nabla_{\epsilon^{\vartheta}} \vartheta(s_d|\epsilon^{\vartheta}), \tag{9}$$

Where $D$ is the training sample batch size, $Q(s,a|\epsilon^Q)$ is the primary critic network with parameter $\epsilon^Q$, $s_d$ is the state of $d$-th sample. The primary critic network is trained by applying a gradient to a loss function, expressed as:

$$L = \frac{1}{D}\sum_{d=1}^{D}(Q(s_d,a_d|\epsilon^Q) - y_d)^2, \quad (10)$$

Where $y_d$ is the target value, which is calculated as:

$$y_d = r_d + \lambda Q^t(s'_d, \vartheta^t(s'_d|\epsilon^{\vartheta^t})|\epsilon^{Q^t}), \quad (11)$$

Where $r_d$ is the reward of the $d$-th sample, $\lambda$ is the discount factor, $Q^t(s,a|\epsilon^{Q^t})$ and $\vartheta^t(s|\epsilon^{\vartheta^t})$ are the target critic and actor networks, with the corresponding parameters $\epsilon^{Q^t}$ and $\epsilon^{\vartheta^t}$, respectively, and $s'_d$ is the next stage of $d$-th sample.

Accordingly, the target networks are updated following a soft-update as:

$$\epsilon^{\vartheta^t} \leftarrow \delta\epsilon^{\vartheta} + (1-\delta)\epsilon^{\vartheta^t},$$
$$\epsilon^{Q^t} \leftarrow \delta\epsilon^{Q} + (1-\delta)\epsilon^{Q^t}, \quad (12)$$

Where $\delta$ is the update coefficient. In training process, the decided action is added with noise to explore the training sample. Therefore, the action decided by DDPG is expressed as:

$$a[t] = \vartheta\big(s[t]|\epsilon^{\vartheta}\big) + \mathcal{N}[t], \quad (13)$$

Where $\mathcal{N}[t]$ is the noise that follows the Ornstein-Uhlenbeck process (Uhlenbeck & Ornstein, 1930).

### *3.3. Deep reinforcement learning framework*

With the DDPG algorithm, the agent can decide the appropriate action at each observation state. However, the decided action may violate the constraints (5b) and (5c). To guarantee the constraints, we propose a two-step action-refined stage. The first step is normalizing the action into a specific range of [0,1]. The second step is defining new power variables $p'_{n,i}, i \in \{1,2\}, n \in \{1,2,..N\}$. To satisfy constraint (5b), the new power variables can be calculated as:

$$p'_{n,1} = p_{n,1}P_n,$$
$$p'_{n,2} = p_{n,2}\big(1 - p'_{n,1}\big)P_n. \quad (14)$$

Besides, the 2-norm of the detection vector $\boldsymbol{w}$ is calculated as:

$$\big||\boldsymbol{w}|\big|_2 = \sqrt{\sum_{m=1}^{M}|w_m|^2}, \quad (15)$$

Where $w_m$ is the $m$-th element of $\boldsymbol{w}$. Therefore, to satisfy the constraint (5c), a new detection vector $\boldsymbol{w}' \in \mathbb{C}^{M\times 1}$ is proposed, where its $m$-th element is calculated as:

$$w'_m = \frac{w_m}{\sqrt{\sum_{m=1}^{M}|w_m|^2}}. \quad (16)$$

As a result, all constraints are satisfied. Consequently, the proposed framework is illustrated in Figure 1. As each time slot $t$, the agent receives the state $s[t]$ from the environment and decides action $a[t]$ using the primary actor-network. The action values are then normalized into a range of [0,1]. Accordingly, the action refined stage is applied, where new power variables $p'_{n,1}[t], p'_{n,2}[t]$ and detection vector $\boldsymbol{w}'$ are calculated as equations (14) and (16), respectively. Then, new actions, including $\pi n,i[t]$, $pn,i'[t]$, and $\boldsymbol{w}'[t]$, are used to interact with the environment.

At each step, a sample combined from $s[t]$, $a[t]$, $r[t]$, and $s'[t]$ is stored in a replay buffer for training the agent, where $s'[t]$ is the next state. In training, the agent randomizes a $D$-size batch of samples from the replay buffer and applies the DDPG algorithm described in sub-section 3.1 to update the networks' parameters at each step.

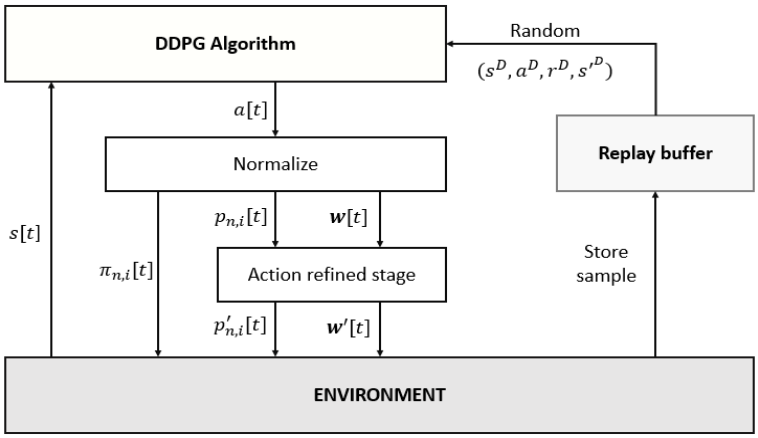


**Figure 1.** Proposed framework

## 4. Result

In this section, an environment established by a BS and 10 UEs is simulated to evaluate the framework's performance. The channel gain vectors randomly follow the Gaussian distribution with a mean of zero and a variance of one. The training convergence of the proposed framework is assessed by training the agent with three learning rate values, where the learning rates of the actor network (ar) and critic network (cr) are selected in (ar, cr) = $\{(1e^{-4}, 1e^{-3}), (1e^{-3}, 1e^{-3}), (5e^{-3}, 5e^{-3})\}$. As illustrated in Figure 2, the algorithm gives the best convergence in case (ar, cr) = $(1e^{-3}, 1e^{-3})$, where it converges after approximately 1,800 episodes, and the reward hits the value of $1.4e^5$, while case $(1e^{-4}, 1e^{-3})$ converges after about 2,800 episodes, and the remaining case gets into the local optimal at the value about $1.28e^5$.

Then, the simulation assesses the system sum rate in different transmit powers and different numbers of BS antennas. As shown in Figure 3, the sum rate increases with the rise of transmit power, which grows about 45% and 65.8% when the transmit power increases from 2dBm to 10dBm in cases M = 4 and M = 8, respectively. Furthermore, more antennas enhance the result, where case M = 8 outperforms case M = 4 in all scenarios.

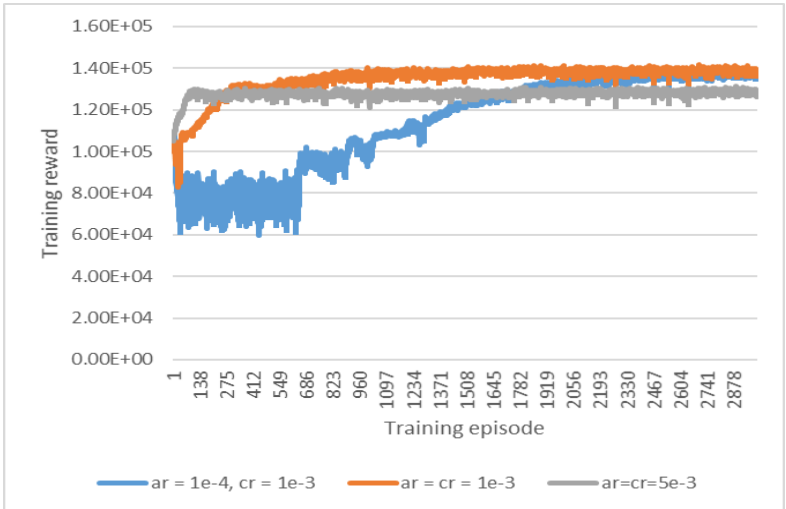


**Figure 2.** Training convergence

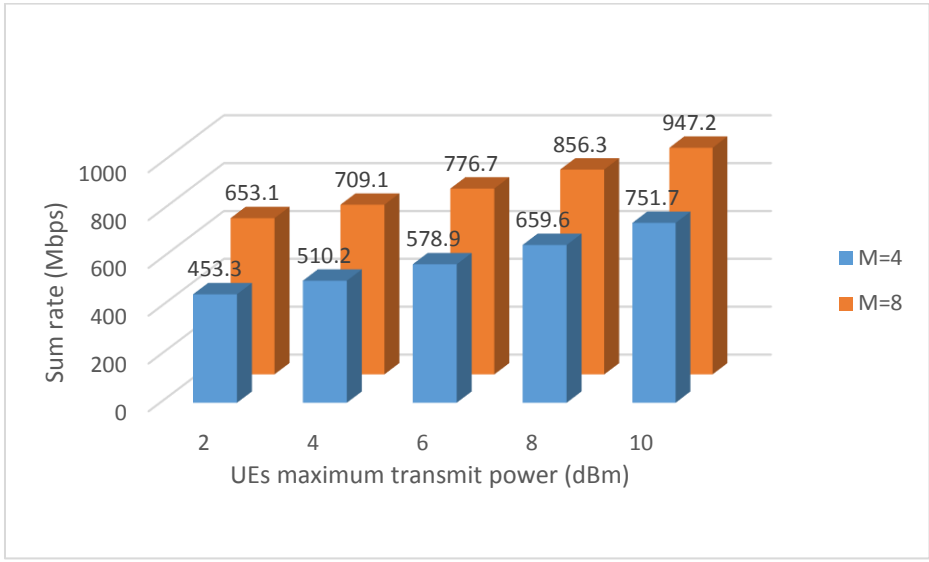


**Figure 3.** Performance in different transmit power

Besides, we evaluate the proposed algorithm by comparing it with local search and random schemes in different communication bandwidths. The local-search approach discretes the action values into discrete values and applies a low-complexity search (Ma et al., 2020) to find the best result at each time slot. The actions are randomly selected from the appropriate range in the random scheme. As illustrated in Figure 4, the proposed algorithm is superior to the remaining schemes, up to 45% and 86% higher than the local search and random schemes, respectively. In addition, high communication bandwidth gives more efficiency in communication, where the sum rate increases with the gain of communication bandwidth.

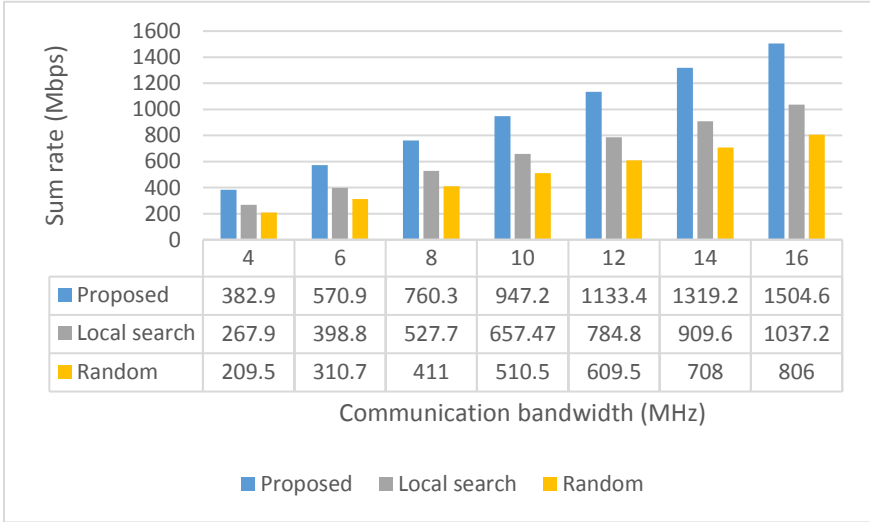


**Figure 4.** Performance in different communication bandwidth

## 5. Conclusions

In this work, we considered an uplink multi-user MIMO system with the aid of the RSMA technique. Here, an optimization problem was established to maximize the sum rate of all UES by considering UEs' transmit powers, decoding order, and detection vector at the BS as variables. To solve the problem, we proposed a DRL framework that employs the DDPG algorithm to train the agent. The simulation results indicated the convergence of the training results with different learning rates. The results also assessed the performance in different scenarios and demonstrated its outperformance compared with some benchmark schemes.

The proposed framework opens many research aspects, such as:

- Considering a complex environment where the channel gains are imperfect knowledge.
- Applying the proposed scheme to mobile edge computing systems.
- Investigating massive MIMO with RSMA.

**References**

de Sena, A. S., Nardelli, P. H. J., da Costa, D. B., Popovski, P., Papadias, C. B., & Debbah, M. (2022). *RSMA for Dual-Polarized massive MIMO networks: A SIC-Free Approach*. Paper presented at the GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., … Wierstra, D. (2015). *Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971*. Paper presented at the ICLR2016 - 6th International Conference on Learning Representations, Vancouver, BC, Canada.

Ma, X., Chen, Z., Chen, W., Li, Z., Chi, Y., Han, C., & Li, S. (2020). Joint channel estimation and data rate maximization for intelligent reflecting surface assisted terahertz MIMO communication systems. *IEEE Access, 8*, 99565-99581. doi:10.1109/ACCESS.2020.2994100

Ma, Y., Ren, S., Quan, Z., & Feng, Z. (2022). Data-driven hybrid beamforming for uplink multi-user MIMO in mobile millimeter-wave systems. *IEEE Transactions on Wireless Communications, 21*(11), 9341-9350. doi:10.1109/TWC.2022.3175878

Mao, Y., Dizdar, O., Clerckx, B., Schober, R., Popovski, P., & Poor, H. V. (2022). Rate-splitting multiple access: Fundamentals, survey, and future research trends. *IEEE Communications Surveys & Tutorials, 24*(4), 2073-2126. doi:10.1109/COMST.2022.3191937

Nguyen, L. V., Vo, Q. T., & Nguyen, T. H. (2023). Adaptive KNN-Based extended collaborative filtering recommendation services. *Big Data and Cognitive Computing, 7*(2), Article 106. doi:10.3390/bdcc7020106

Nguyen, T. H., & Park, L. (2023). HAP-Assisted RSMA-Enabled vehicular edge computing: A DRL-based optimization framework. *Mathematics 11*(10), Article 2376. doi:10.3390/math11102376

Nguyen, T. H., Park, H., & Park, L. (2023). *Recent studies on deep reinforcement learning in RIS-UAV communication networks*. Paper presented at the 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia.

Nguyen, T. H., Park, H., Seol, K., So, S., & Park, L. (2023). *Applications of deep learning and deep reinforcement learning in 6G networks*. Paper presented at the 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France.

Park, J., Choi, J., Lee, N., Shin, W., & Poor, H. V. (2023). Rate-Splitting multiple access for downlink MIMO: A generalized power iteration approach. *IEEE Transactions on Wireless Communications, 22*(3), 1588-1603. doi:10.1109/TWC.2022.3205480

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tran, S. V. T., Bao, Q. L., Nguyen, T. L., & Park, C. (2022). *A framework for computer vision-aided construction safety monitoring using collaborative 4D BIM*. Paper presented at the ICCEPM 2022 - The 9th International Conference on Construction Engineering and Project Management, Las Vegas, NV, USA.

Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical Review*, *36*(5), Article 823.

Yang, Z., Chen, M., Saad, W., Xu, W., & Shikh-Bahaei, M. (2022). Sum-Rate maximization of uplink Rate Splitting Multiple Access (RSMA) communication. *IEEE Transactions on Mobile Computing*, *21*(7), 2596-2609. doi:10.1109/TMC.2020.3037374

Zheng, G., Wong, K. K., & Ng, T. S. (2009). Energy-efficient multiuser SIMO: Achieving probabilistic robustness with gaussian channel uncertainty. *IEEE Transactions on Communications*, *57*(6), 1866-1878. doi:10.1109/TCOMM.2009.06.070574