A Review of Proximal Policy Optimization for Uplink Multi-user SIMO-RSMA Systems

Huy Dang Mac, Kiet Nguyen Tuan Tran, Dongwook Won, and Sungrae Cho
School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
Email: {hdmac, knttran, dwwon}@uclab.re.kr, srcho@cau.ac.kr

***Abstract*—Rate-Splitting Multiple Access (RSMA) has emerged as a powerful multiple access technique for interference management in next-generation wireless networks. In uplink multi-user single-input multiple-output (SIMO) systems, RSMA enables a base station to decode multi-user transmissions more flexibly by splitting messages and partially decoding interference. Optimizing such RSMA systems is challenging due to the complex resource allocation and decoding order decisions required, especially under dynamic channel conditions. Recently, deep reinforcement learning (DRL) approaches, particularly Proximal Policy Optimization (PPO), have been applied to efficiently learn near-optimal policies for uplink RSMA resource allocation. This paper provides a comprehensive review of PPO-based techniques for uplink multi-user SIMO-RSMA systems. We outline the fundamentals of SIMO-RSMA, describe the PPO approach for solving the sum-rate and fairness optimization problems in these systems, and survey state-of-the-art PPO-based solutions from 2022 to 2025. A comparative analysis of recent studies is presented, and open challenges as well as future research directions are discussed.**

***Index Terms*—Rate-splitting multiple access, uplink multi-user SIMO, proximal policy optimization, deep reinforcement learning, resource allocation.**

I. Introduction and Motivation

Rate-splitting multiple access (RSMA) is a flexible multi-antenna transmission strategy that splits user messages into common and private parts to manage multi-user interference [1], [2]. Originally developed for downlink multi-antenna systems, RSMA has demonstrated the ability to bridge and outperform conventional space-division multiple access (SDMA) and non-orthogonal multiple access (NOMA) techniques [1]. The core idea is that by allowing part of the interference to be decoded and canceled while treating the remaining interference as noise, RSMA generalizes existing multiple access schemes and improves network performance [2].

While most early RSMA research focused on downlink transmissions, the RSMA principle can also be applied to uplink scenarios. In an uplink multi-user single-input multiple-output (SIMO) system, multiple single-antenna users transmit simultaneously to a base station (BS) equipped with multiple antennas [3]. By employing RSMA reception strategies at the BS—such as decoding a portion of interfering signals via successive interference cancellation (SIC) and treating the rest as noise—the uplink interference can be more effectively managed [4]. This enhanced interference management translates to higher throughput and robustness. For instance, uplink RSMA has been shown to achieve significant sum-rate gains (on the order of 10–80%) over baseline NOMA and orthogonal access schemes in various conditions [4]. However, to fully realize these benefits, one must solve a challenging joint optimization of power allocation, receive beamforming, and decoding order selection for the RSMA uplink system [3]. Traditional optimization methods for this problem often rely on idealized assumptions (e.g., perfect channel state information) and convex approximations, which may struggle under practical dynamics and uncertainties [3].

In this context, deep reinforcement learning (DRL) has emerged as a promising approach to handle complex, dynamic optimization in wireless networks [5]. DRL enables an agent to learn good decision policies (e.g., how to allocate power or adjust decoding order in response to channel changes) through interactions with the environment, without requiring an explicit model of the wireless channel or traffic dynamics. Among DRL algorithms, Proximal Policy Optimization (PPO) is particularly appealing due to its training stability and ability to handle continuous action spaces [5]. PPO uses an actor-critic framework and a clipped surrogate objective to ensure policy updates do not deviate drastically, which makes it effective for resource allocation problems. Several recent works have applied PPO or PPO-based methods to multi-user communication scenarios and demonstrated superior performance compared to conventional optimization or heuristic schemes [5], [6], [8], [9]. These successes motivate a deeper review of how PPO can be leveraged specifically for uplink multi-user SIMO-RSMA systems.

In this paper, we aim to review and synthesize the state-of-the-art in applying PPO to uplink multi-user SIMO-RSMA systems. We provide background on the SIMO-RSMA model and the PPO algorithm, discuss how PPO is utilized to learn resource allocation policies in this context, and compare recent studies (2022–2025) that employ PPO in related RSMA optimization problems. We also identify key challenges (e.g., scalability, hybrid action spaces, sample efficiency) in deploying PPO for these systems and outline potential directions for future research.

The remainder of this paper is organized as follows. Section II presents the background on uplink SIMO-RSMA systems and relevant technologies. Section III details the PPO-based approach in SIMO-RSMA and reviews representative recent works, with a comparative summary provided in Table I. Section IV discusses open challenges and future research directions. Finally, Section V concludes the paper.

II. Background and Technologies

A. Uplink Multi-user SIMO-RSMA Systems

An uplink multi-user SIMO-RSMA system consists of K single-antenna users simultaneously transmitting to a BS with M receive antennas. All users share the same time-frequency resources, resulting in co-channel interference at the BS. The BS employs RSMA-based reception to manage this interference. In essence, the BS can decode part of the combined signals as a *common* stream that captures a portion of the interference, while treating the remainder of the interference as noise when decoding each user's *private* stream. By performing successive interference cancellation, the BS first decodes the common stream (which is a function of multiple users' signals), subtracts it from the aggregate received signal, and then decodes the individual private streams of each user [1], [2]. This process generalizes the uplink decoding: if the common stream is configured to carry no information, RSMA reduces to treating all interference as noise (as in orthogonal multiple access), whereas if the common stream carries all interfering content, it resembles decoding interference fully (as in power-domain NOMA). A properly chosen common/private splitting strategy allows intermediate operation, yielding robust performance under a variety of channel conditions [2].

To implement uplink RSMA, the system must determine several aspects: (1) each user's transmit power allocation (potentially splitting power between common and private portions of its message), (2) the decoding order of streams at the BS (which stream to decode first, second, etc.), and (3) the receive beamforming vectors at the BS for each stream, which leverage the M antennas to separate the signals [3]. The overall objective is often to maximize a network utility (e.g., the sum of user achievable rates or a fairness metric) subject to power constraints and decoding feasibility. This is a challenging non-convex problem because the decisions are interdependent; for example, the optimal power allocation depends on the decoding order and vice versa. Conventional approaches have applied iterative optimization, such as alternating between optimizing beamformers and power splits, or using techniques like successive convex approximation to handle the non-convex rate expressions [3], [4]. While such methods can reach a locally optimal solution, they require accurate channel state information and can be computationally intensive, making them less adaptable to fast-changing environments.

B. Deep Reinforcement Learning and PPO

Deep reinforcement learning offers an alternative by modeling the resource allocation problem as a Markov decision process (MDP) and learning a control policy through trial-and-error. In an uplink RSMA context, the DRL agent (located at the BS or network controller) can observe the state (e.g., channel gains of all users, queue lengths, interference levels) and then take actions (such as setting users' power levels, selecting a decoding order, or adjusting beamforming) at each time slot. A well-designed reward function (for instance, the sum-rate achieved or a weighted sum-rate reflecting QoS priorities) guides the agent to improve performance over time. Notably, DRL does not require explicit models of the channel or interference; it learns directly from feedback (reward signals) by interacting with the environment, which is advantageous under uncertainty or complexity that defies analytical solutions.

Among DRL algorithms, Proximal Policy Optimization (PPO) has gained popularity due to its balance of implementation simplicity and reliable convergence [11], [12]. PPO is a policy gradient method that uses an actor-critic architecture: an actor network outputs the policy (probability distribution over actions, or direct action values in continuous space) and a critic network estimates the value function (expected return) to help compute advantage estimates. PPO improves training stability by limiting how much the policy can change at each update. It does so by formulating a clipped surrogate objective that penalizes large deviations from the previous policy during gradient updates [11], [13]. This mechanism prevents unstable swings in the policy and has been shown to yield stable learning in many continuous control problems. In wireless communications, PPO is well-suited for problems like power control or beamforming where the action space can be continuous and high-dimensional [5]. Furthermore, PPO inherently handles exploration vs. exploitation trade-offs and can be combined with reward shaping or constraint handling techniques (e.g., penalty terms) to enforce power or interference constraints.

Recent studies have started to apply PPO to RSMA-related optimization tasks. For example, in a downlink RSMA power allocation problem with unknown channel dynamics, PPO was used to learn the transmit power policy that maximizes sum-rate, outperforming baseline algorithms that assumed partial channel knowledge [5]. In another work, a PPO-based approach was employed to jointly optimize resource allocation in a satellite communication system using RSMA, adapting to time-varying channel conditions in low Earth orbit links [6]. These applications demonstrate that PPO can successfully handle the coupling of decisions and uncertainties inherent in RSMA systems. In the next section, we review in detail how PPO has been applied specifically to RSMA scenarios, including uplink multi-user SIMO contexts and related systems, and compare key achievements of recent works.

III. PPO Approach in SIMO-RSMA System

PPO-based solutions for RSMA systems involve defining the agent's observations, actions, and rewards to capture the RSMA resource allocation problem. Typically, the state observed by the PPO agent includes channel information for all users (e.g., channel gains or estimates for each user's link to the BS) and possibly other context like the users' buffer statuses or interference levels. The action can be a vector consisting of each user's power allocation (and potentially their rate-split ratios if the user transmits a common and private part) as well as discrete choices like decoding order. In cases where decoding order needs to be optimized, one approach is to incorporate it into the action space by

enumerating possible orders or to handle it with a separate mechanism or agent, as the action space would otherwise be hybrid discrete-continuous. The reward is often chosen as the sum-rate achieved by all users under the chosen allocation and decoding strategy, or a utility that reflects fairness (e.g., minimum rate among users for max-min fairness). The agent then learns a policy that maps states to actions to maximize the expected cumulative reward (long-term performance).

One of the first studies to apply PPO in an RSMA context was [5], which considered a two-user downlink RSMA system with an unknown channel model. The PPO agent was used to determine the optimal power allocation for the common and private messages to maximize the sum-rate. This work demonstrated that the learned policy could outperform traditional solutions that rely on perfect or statistical CSI, especially when the channel was dynamically varying. Another related work [7] extended this idea to include covert communication constraints (where one user's transmission must remain undetected by a malicious receiver) in an RSMA network. The PPO algorithm was utilized to jointly optimize power allocation and rate control, achieving a balance between spectral efficiency and covertness.

In uplink settings, *deep deterministic policy gradient* (DDPG) algorithms have been explored for RSMA as well. For example, a DDPG-based approach tackled uplink multi-user SIMO-RSMA sum-rate maximization by learning the users' power control and the BS's decoding order. PPO, being an on-policy algorithm, offers some different advantages in this context, such as improved stability in training at the cost of more sample usage compared to off-policy methods like DDPG. While specific literature on PPO for uplink RSMA is still limited, the methods and findings from downlink and related scenarios can be translated to uplink. The BS in uplink can serve as the agent that learns how to optimally decode and allocate power (potentially by sending power control commands to users or assuming users adjust their power based on BS's decisions). The successful application of PPO in downlink RSMA and other multi-user interference problems suggests that PPO is a viable approach for the uplink case as well.

Table I summarizes several representative works from 2021 to 2024 that employed PPO-based algorithms for RSMA or closely related multi-user communication systems. The table outlines the scenario, objective, DRL approach, and key results of each study. As shown, PPO has been used in a variety of contexts—including terrestrial cellular, satellite communications, and networks aided by reconfigurable intelligent surfaces (RIS)—and for objectives ranging from sum-rate maximization to energy efficiency and fairness. These studies consistently report that PPO-based solutions can approach or exceed the performance of conventional optimized schemes, especially in dynamic environments where classical methods struggle.

Overall, the reviewed works illustrate that PPO-based DRL agents can effectively learn resource allocation strategies in complex multi-user RSMA systems. They can adapt to different channel conditions (radio frequency or optical wireless), network architectures (with or without intelligent surfaces or relays), and performance goals. A common theme is that by learning directly from the environment, PPO can exploit the structure of the interference and channel dynamics that might be intractable to model in closed-form, thus finding efficient solutions that static or one-shot optimization methods might miss.

IV. Challenges and Future Works

Despite the promising results of PPO in RSMA systems, several challenges remain to be addressed to fully leverage this approach in practical multi-user uplink scenarios:

Scalability and High-Dimensional Spaces: As the number of users K or antennas M grows, the state and action spaces for the PPO agent increase significantly. A large state space (e.g., full channel gain matrices for many users) can slow down learning and require extensive training data. Similarly, a high-dimensional continuous action (like a power level for each user and possibly each stream) makes the policy network more complex. Future research could explore dimensionality reduction techniques or multi-agent formulations (splitting the optimization across multiple agents) to maintain scalability. Hierarchical RL is another avenue, where one agent might set high-level parameters (like decoding order or grouping of users) and another fine-tunes power allocations.

Hybrid Discrete-Continuous Decisions: Uplink RSMA optimization involves discrete choices (decoding order selection, user scheduling in some cases) alongside continuous ones (power or beamforming vectors). Standard PPO is designed for either continuous or discrete action spaces, but not both simultaneously. While one could discretize continuous actions or use separate neural network outputs for discrete decisions, these approaches may not be efficient. There is a need for advanced methods to handle hybrid action spaces. One potential direction is to combine PPO with search or combinatorial algorithms for the discrete part—for example, using PPO to optimize power for a given decoding order and employing a secondary search (or another RL agent) to find the best order. Some recent works integrated convex optimization steps into the learning loop (as in PPO-SCF [9]) to guarantee constraint satisfaction, which could be extended to handle discrete decisions as well.

Sample Efficiency and Training Overhead: PPO, being an on-policy algorithm, often requires a large number of interactions with the environment to converge to an optimal policy. In a simulated environment, this translates to many channel realizations and evaluations of the reward, which can be time-consuming. Techniques to improve sample efficiency, such as experience replay (though not straightforward in on-policy methods), meta-learning to quickly adapt policies to new scenarios, or model-based approaches that learn an approximate environment model for planning, are worth investigating. Transfer learning could also be valuable: a policy trained for a certain range of network conditions (user count,

TABLE I
REPRESENTATIVE PPO-BASED STUDIES ON RSMA RESOURCE OPTIMIZATION (2021–2024)

Work (Year)	Scenario and Objective	Approach and Key Findings
Nguyen *et al.* [5] (2021)	Downlink two-user MISO RSMA; maximize sum-rate under unknown channel model	PPO algorithm for power allocation (common vs private stream power). Achieved ∼10% higher sum-rate than baseline with imperfect CSIT, demonstrating RSMA gains without explicit channel model.
Huang *et al.* [6] (2022)	Downlink LEO satellite MISO RSMA; maximize sum-rate (power allocation in satellite downlink)	Deep RL (PPO-based) resource allocation adapting to fast time-varying LEO channels. Outperformed static optimization by adjusting transmit power in real-time, improving throughput in 6G satellite links.
Nguyen *et al.* [7] (2023)	Downlink RSMA with covert communications; joint power allocation and rate control	PPO-based policy optimizing spectral efficiency while maintaining covertness. Showed that RSMA with learned policy met covert constraints and achieved higher sum-rate than heuristic schemes.
Meng *et al.* [8] (2024)	STAR-RIS assisted RSMA (downlink); maximize sum-rate for users	PPO-based joint optimization of transmit power and reflecting surface configuration. Demonstrated fast convergence and notable sum-rate gain over baseline algorithms (e.g., up to 20% higher sum-rate than non-RIS or non-PPO schemes).
Zhang *et al.* [9] (2024)	RIS-aided MISO RSMA; maximize secrecy energy efficiency (SEE)	Proposed PPO-SCF (PPO with successive convexification) to handle continuous phase shifts and power allocation under secrecy constraints. Achieved higher SEE compared to conventional optimization, highlighting PPO's ability to handle physical layer security and efficiency trade-off.
Guo *et al.* [10] (2024)	VLC (Visible Light) downlink MISO with RSMA; maximize secrecy energy efficiency	Developed a DS-PPO (Dual Stage PPO) approach to jointly optimize beamforming and power in an IRS-assisted visible light RSMA system. Results showed improved worst-case (min-user) rate and energy efficiency relative to baseline PPO, thanks to the specialized training strategy.

mobility, etc.) could be fine-tuned for a new scenario rather than training from scratch.

Robustness and Generalization: Wireless environments are highly variable. A PPO agent trained under certain assumptions (e.g., a particular distribution of channel conditions or number of active users) might perform sub-optimally if those conditions change. Ensuring that the learned policy generalizes beyond the training scenarios is crucial for real deployment. This may involve training the agent across a wide variety of random environments (domain randomization) or incorporating robustness in the objective (e.g., optimizing worst-case performance). Another future direction is safe reinforcement learning, which ensures the agent respects critical constraints (like not exceeding power limits or causing outage for users) throughout the learning process, not just at convergence.

Computational and Deployment Challenges: Implementing a DRL solution like PPO in a live network poses practical challenges. The inference delay (to run the neural network and output an action) must be small enough for real-time control in the uplink (which may have scheduling intervals on the order of milliseconds). Custom hardware (ASICs or FPGAs) for neural network acceleration at the BS, or cloud-assisted control, might be necessary. Additionally, the training will likely occur offline using a simulator; discrepancies between the simulated model and reality (model mismatch) can degrade performance. Developing methods for online fine-tuning or continual learning can help adapt the policy to actual network measurements.

In summary, future work should focus on making PPO-based RSMA solutions more robust, efficient, and scalable. Integrating domain knowledge (e.g., using known optimal strategies for simple cases to guide training or initialize the policy) and combining learning with traditional optimization (to enforce hard constraints or provide initial feasible solutions) are promising avenues. As the wireless community moves towards 6G and beyond, where network management will heavily involve AI techniques, addressing these challenges will be key to operationalizing DRL approaches like PPO for multi-user RSMA and other advanced multiple access schemes.

V. CONCLUSION

This paper reviewed the application of Proximal Policy Optimization for uplink multi-user SIMO-RSMA systems. RSMA offers a powerful framework for managing interference by flexibly splitting and decoding messages, but optimizing its operation in uplink multi-antenna scenarios is complex. DRL techniques, and PPO in particular, have shown great potential in tackling these challenges by learning resource allocation policies that adapt to changing conditions. We surveyed recent PPO-based approaches in both uplink and downlink RSMA contexts, highlighting that PPO agents can achieve significant gains in sum-rate, energy efficiency, and fairness over traditional methods. We also discussed several open challenges, including scalability to larger networks, handling hybrid action spaces, ensuring sample-efficient and robust learning, and deploying such solutions in real systems. Addressing these issues will be crucial for the future integration of learning-driven optimization in wireless networks. Nonetheless, the advancements to date indicate that PPO will play an important role in enabling intelligent, high-performance multiple access strategies in 6G and beyond.

VI. Acknowledgment

This work was supported in part by the IITP (Institute of Information & Communications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) (IITP-2026-RS-2022-00156353, 50%) grant funded by the Korea government (Ministry of Science and ICT) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00209125).

References

[1] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-Splitting Multiple Access for Downlink Communication Systems: Bridging, Generalizing, and Outperforming SDMA and NOMA," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 133, 2018.

[2] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-Splitting Multiple Access: Fundamentals, Survey, and Future Research Trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

[3] Z. Yang, M. Chen, W. Saad, X. Wei, and M. Shikh-Bahaei, "Sum-Rate Maximization of Uplink Rate-Splitting Multiple Access (RSMA) Communication," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2596–2609, 2022.

[4] M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Rate Splitting Multiple Access for Sum-Rate Maximization in IRS-Aided Uplink Communications," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2246–2261, 2023.

[5] N. Q. Hieu, D. T. Hoang, D. Niyato, and D. I. Kim, "Optimal Power Allocation for Rate Splitting Communications with Deep Reinforcement Learning," *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2820–2823, Dec. 2021.

[6] J. Huang, Z. Feng, K. Xiong, W. Wei, P. Fan, and K. B. Letaief, "Deep Reinforcement Learning-Based Power Allocation for Rate-Splitting Multiple Access in 6G LEO Satellite Communication System," *IEEE Wireless Communications Letters*, vol. 11, no. 10, pp. 2185–2189, Oct. 2022.

[7] N. Q. Hieu, D. T. Hoang, Q.-D. Ho, D. Niyato, and D. I. Kim, "Joint Power Allocation and Rate Control for RSMA Networks with Covert Communications," *IEEE Transactions on Communications*, vol. 71, no. 4, pp. 2274–2287, Apr. 2023.

[8] C. Meng, K. Xiong, W. Wei, and K. B. Letaief, "Sum-Rate Maximization in STAR-RIS-Assisted RSMA Networks: A PPO-Based Algorithm," *IEEE Internet of Things Journal*, vol. 11, no. 4, pp. 3029–3043, Feb. 2024.

[9] W. Zhang, K. Xiong, R. Zhang, P. Fan, and K. B. Letaief, "SEE Maximization in RIS-Aided Network with RSMA: A PPO-SCF Method," *IEEE Wireless Communications Letters*, vol. 13, no. 12, pp. 3315–3319, Dec. 2024.

[10] Y. Guo, J. Hou, J. Zhang, J. Shi, J. Ye, and X. Sun, "Secrecy Energy Efficiency Maximization in IRS-Assisted VLC MISO Networks with RSMA: A DS-PPO Approach," *arXiv preprint arXiv:2411.09146*, Nov. 2024.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] Q. Yang and S.-J. Yoo, "Multi-objective task offloading optimization using deep reinforcement learning with resource distribution clustering," *ICT Express*, vol. 11, no. 4, pp. 734–742, 2025.

[13] J. A. Bermúdez, P. Morales, H. Pempelfort, M. Araya, and N. Jara, "Understanding deep reinforcement learning: Enhancing explainable decision-making in optical networks," *ICT Express*, vol. 11, no. 5, pp. 969–973, 2025.

[14] W. J. Yun *et al.*, "Cooperative Multiagent Deep Reinforcement Learning for Reliable Surveillance via Autonomous Multi-UAV Control," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7086–7096, Oct. 2022.

[15] W. J. Yun, D. Kwon, M. Choi, J. Kim, G. Caire, and A. F. Molisch, "Quality-Aware Deep Reinforcement Learning for Streaming in Infrastructure-Assisted Connected Vehicles," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2002–2017, Feb. 2022.