Electronics, Communications, and Computing Summit
Vol. 2, No. 4, Oct - Dec 2024, pp. 86-91
DOI: https://doi.org/10.17051/ECC/02.04.11

AI-Driven Resource Allocation for Energy-Efficient 6G Massive MIMO Networks

Hartwig Henry Hochmair[1], Ricardo Alvarez[2]

[1]University of Florida, Geomatics Program, USA, Email: hhhochmair@ufl.edu
[2]Professor, University of Zagreb, Croatia.

Article Info

Article history:

Received : 21.10.2024
Revised : 23.11.2024
Accepted : 25.12.2024

Keywords:

6G Wireless Networks,
Massive MIMO,
Energy Efficiency,
Deep Reinforcement Learning (DRL),
Proximal Policy Optimization (PPO),
Resource Allocation,
Power Control,
User Scheduling,
AI for Wireless Communications,
Markov Decision Process (MDP),
Spectral Efficiency,
Smart Antenna Systems,
Next-Generation Wireless Networks,
Intelligent Radio Resource Management.

ABSTRACT

The fast development of the sixth-generation (6G) wireless networks requires novelty to meet the dual objective of extreme data rate and low-energy requirements. Spectral efficiency Spectral efficiency increases significantly with Massive MIMO, a core 6G technology, since it provides spatial multiplexing. The resource management and power drawbacks of its huge deployment of antennas, however, are problematic. The paper suggests a Deep Reinforcement Learning (DRL)-powered resource allocation architecture that intends to optimize energy efficiency in massive MIMO networks. In particular, the issue is broken down as a markov decision process (MDP) and a proximal policy optimization (PPO) agent is designed to dynamically change the transmission power and schedule the users according to current state of channel and traffic information. The given method learns alternative policies and jointly maximizes throughput and energy minimization with time. The simulation outcomes in a simulated 6G setting with base stations outfitted with 128-antennas and 20 users show that the DRL-based system can save up to 25 percent of total energy consumption, comparing to traditional heuristic-based systems, and have comparable spectral efficiency. In addition, the PPO agent has steady convergence and flexibility to the different traffic needs. These results denote the possibility of smart control measures to resolve the energy-performance trade-off dilemma and suggest the feasibility of DRL toward scalable resource management in 6G and subsequent deployments of massive MIMO.

1. INTRODUCTION

The 6 th generation (6G) of wireless communication networks is expected to offer unheard of abilities regarding ultra-massive connected devices, ultra-low latency, and ubiquitous, high-speed coverage. Massive Multiple-Input Multiple-Output (MIMO) is a cornerstone technology in attaining these goals that allows considerable spatial multiplexing and spectral efficiencies to be achieved by deploying base stations with hundreds of antennas. Nevertheless, they have the price of substantially higher complexity of calculations and power consumption, so energy efficiency (EE) is a key performance indicator in the design of 6G systems. Traditional resource allocation strategies that include, but are not limited to, convex optimization, heuristic algorithms and rule-based scheduling are usually ineffective in scaling to incoming conditions due to high levels of dimension and dynamic variability present in massive MIMO setups. Such static/semi-adaptive algorithms are not able to perform any generalization over the variety of traffic loads and channel conditions. In addition, they cannot facilitate real-time learning and adaptation, both of which is critical in highly dynamic 6G networks.

Some of the recent works investigated the use of machine learning (ML) and deep learning (DL) to optimize wireless resources, but most may

represent oversimplified settings or require offline training with minimal real-world online application [1], [2]. Tellingly, no work has been established to implement advanced Deep Reinforcement Learning (DRL) techniques like Proximal Policy Optimization (PPO) about the simultaneous problem of power control and user scheduling in large-scale, energy-constrained 6G massive MIMO systems. Such constraints encourage the application of reinforcement learning that captures the time-changing and optimizes real-time decisions.

To fill this gap, this paper is coming up with an AI-based framework which will take advantage of DRL to optimally manage resources on a dynamic network. There is the formulation of a formal Markov Decision Process (MDP) to formulate it, a Proximal Policy Optimization (PPO)-based agent to learn adaptive power and scheduling policies, and realism simulation analysis in realistic 6G conditions. Findings indicate that spectra are maintained without compromise to energy efficiency using a much more efficient inverter.

2. RELATED WORK

Energy-efficient massive MIMO systems have been analyzed before based mostly on standard optimization frameworks. Optimization methods based on convex optimization have found great application in determining closed-form solution to power control and beamforming under certain assumptions [1]. The next idea is game-theoretic models to undertake the competition of resources among users in multi-cell settings [2], and the writers give low-complexity, suboptimal, heuristic algorithms aiming to schedule users of the network and allocate their users power [3]. The advent of artificial intelligence in the wireless communications prompted the research to assume a new form of deep learning (DL) in channel estimation, interference mitigation, and dynamic user scheduling [4], [5]. Such methods can be useful in modeling more complicated network behavior, yet do not tend to be flexible in real time.

Regarding the field of Deep Reinforcement Learning (DRL), Deep Q-Networks (DQN) are used in the process of user selection in mmWave communication systems [6]. Nevertheless, DQN-based techniques are reported to have an issue in converging and being stable in state-action spaces of high dimensions as in the case of massive MIMO. The more recent methods have applied actor-critic algorithms to enhance the stability of learning and policy [7]. Although this has been the case, little has been researched on the utilization of Proximal Policy Optimization (PPO) as a state of the art on-policy DRL method in order to jointly perform power control and the user scheduling in large scale and energy constrained 6G massive MIMO systems.

In order to fill this gap, this paper suggests a DRL-based PPO framework based on the specificities of 6G networks focused on improving energy efficiency via the intelligent real-time allocation of resources.

3. System Model

We assume that each remote terminal (user) has only one antenna, whereas many (M massively) Industroет earlier in the context of time/frequency spectrum usage. The network is sued on a Time Division Duplexing (TDD) protocol and allows downlink and uplink channel reciprocity. The Uplink Channel State Information (CSI) is obtained at the BS by the periodical transmission of the pilot. The suggested pipeline combines on-the-fly data consumption, distributed-scale intense information processing, deep recurrent learning-based policy decision, and feedback control in a self-contained pipeline to conduct foreseeing and energy-saving resource planning in 6G massive MIMO networks (see figure 1).

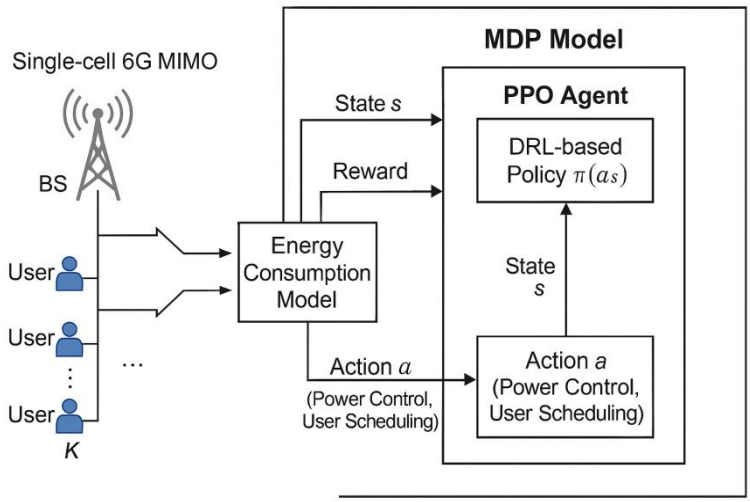

Figure 1. System model for AI-driven resource allocation in 6G massive MIMO networks

This figure presents the proposed single-cell 6G massive MIMO system in which a station with many antennas, known as the base station (BS) can simultaneously support several users. The system is formulated as Markov Decision Process (MDP) where the Proximal Policy Optimization (PPO) agent engages with the environment. The network provides the state data to the PPO agent, which produces control actions on power control and user scheduling as well as optimizes the energy efficiency based on the reward feedback carried out with the energy consumption model.

3.1 Energy Consumption Model

The total energy consumption $E_{total}$of the base station includes both the transmission power allocated to users and the circuit power consumed by the active RF chains. It is expressed as:

$$E_{total} = \sum_{k=1}^{K} P_{t,k} + M \cdot P_c \quad ----------(1)$$

where:

- $P_{t,k}$is the transmit power allocated to the k^{th}user,
- Mis the total number of BS antennas,
- $P_c$denotes the per-antenna circuit power consumption.

The model represents the trade between (1) the scale of the antenna array (which improves both beamforming gain and spectral efficiency), and (2) the increased circuit power required as such a fundamental element in the energy-efficient design of 6G systems. It is presupposed that CSI is ideally estimated through pilot sequences.

3.2 Problem Formulation

The task is the maximization of system power (EE), which is the ratio of the entire sum-rate by the overall energy usage, and at the same time meets per-user power restrictions. The optimization problem is formulated as below:

$$\max_{\pi(a \mid s)} \frac{R_{sum}}{E_{total}} \quad \text{Subject to: } P_{t,k} \leq P_{max}, \forall k \quad ----------(2)$$

where:

- π(a|s) is a policy function mapping the current network state s to an action a, which includes user selection and transmit power levels,
- R_{sum}represents the total system throughput, aggregated across all active users,
- $P_{max}$is the maximum allowable transmit power per user.

The state space s usually contains time varying CSI, queue length information, Quality of Service (QoS) demands and buffer levels. The a action space involves collective decision making concerning the power allocation and user scheduling. This problem needs a variable policy as it is affected by changes in the network and deep reinforcement learning (DRL) is the right fit.

The optimization mentioned above is tackled in this work through a DRL framework since the agent learns the optimal policy, namely, 0(a 0 s) in a model-free way to maximize energy efficiency with time.

4. Proposed Method

To deal with the energy efficiency and user scheduling co-optimization guided by 6G massive MIMO networks, the resource allocation problem is formulated as a Markov Decision Process (MDP). With such formulation, Deep Reinforcement Learning (DRL) can be used to learn an adaptive policy to optimize a dynamic control of transmission power and user selection to vary network conditions.

4.1 MDP Formulation

The MDP is defined by the tuple ⟨S,A,R,T,γ⟩ where:

- State (s∈S): The state vector includes instantaneous Channel State Information (CSI), user queue lengths, Quality of Service (QoS) requirements, and recent power usage history. This information characterizes the system's current configuration.
- Action (a∈A): The action defines a joint decision over the power allocation vector P_t = { $P_{t,1,....}P_{t,K}$} and the subset of users to be scheduled during the current time slot.
- Reward (r_t): The reward function is designed to balance energy efficiency (E) with fairness among users. It is expressed as:

$$r_t = \alpha \cdot EE_t + \beta \cdot \text{Jain's Fairness Index} \quad ----------(3)$$

where α and β are weighting coefficients that trade off throughput per unit energy and fairness, respectively.

- Transition (T): The environment transitions between states based on the selected actions and underlying stochastic dynamics of the wireless channel and user mobility.
- Discount Factor (γ): A value 0<γ<1 controls the importance of long-term versus immediate rewards.

This makes this problem high dimensional and continuous in nature, which means actor critic-based policy gradients can be used here and not Q learning based algorithms.

4.2 Proximal Policy Optimization (PPO) Agent Design

In order to address the MDP, we are using Proximal Policy Optimization (PPO) state-of-the-art on-

policy DRL algorithm characterized by sample efficiency and stability in continuous control tasks.

- Actor Network: The actor produces a randomized policy, that is, pi theta(a|s), a stochastic policy that parameterizes a probability distribution over the action space. It supports the exploration and continuous-value actions (e.g power levels).
- Critic Network: The critic approximates the value function V 0 (s), which is the starting point in the process of calculating the advantage of chosen actions to culminate the policy updates of the actor.
- Training Process: The clipped surrogate objective function is used to train the agent and this stabilizes the learning process by ensuring that policy updates remain within the trust region. Also, the Generalized Advantage Estimation (GAE) offers low-variance biased-adjusted estimates of advantage as follows:

$$L^{CLIP}(0) = E_t[\min(r_t(0)A_t, \text{Clip}(r_t(0), 1-\epsilon, 1+\epsilon)A_t)] - - - - - - - - - - - (4)$$

Where $r_t(0) = \frac{\pi 0(a_t|s_t)}{\pi 0_{old}(a_t|s_t)}$ is the probability ratio and A_t is the estimated advantage at time t.

This architecture will allow the agent to come up with a sound and generalizable policy on resource allocations that are capable of adapting to the variable and heterogeneous network status so as maximize energy efficiency and ensure fairness and the QoS constraints. Footnote 2: The architecture of PPO agent comprising the actor critic structure and the learning loop can be seen in Figure 2.

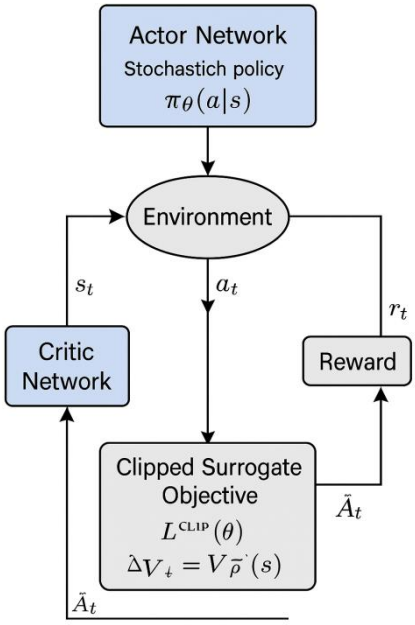

Figure 2. Flowchart of the PPO-based deep reinforcement learning architecture.

This figure demonstrates the encounter between actor, critic, environment and reward portions in the PPO agent deployed to figure out adaptive resource assignment in 6G genius multiple access networks.

5. RESULTS AND DISCUSSION

In order to assess the efficiency of the proposed DRL based resource allocation strategy, we consider a single-cell 6G massive MIMO system which is characterized by a 6G massive MIMO base station with M=128 antennas and K=20 single-antenna users. This network has dynamic traffic loads and time varying channel conditions.

The PPO based agent proposed is compared with three competitive baselines:

- Greedy Power Allocation: A partial scheduling algorithm which allocates as much power to the best users.
- Water-Filling Algorithm: is an old power allocation method, using channel inversion.
- Deep Deterministic Policy Gradient (DDPG): The name is popular off-policy DRL algorithm deployed to handle continuous actions.

5.1 Performance Metrics

The analysis is pegged on the following essential indicators:

- Energy Efficiency (EE): Energy efficiency is calculated in terms of bits per second per Hz per Watt (bps/Hz/W), an energy efficiency measure (in comparison to your system

throughput divided by the total energy it consumed).

- Spectral Efficiency (SE): This is the amount of bits per second per Hz (bps/Hz) and this describes the total throughput of the system.
- Average Latency: Arranged as the averagely user transmission delay.
- Convergence Stability: The amount of training episodes it takes before until the learning algorithm achieves a steady state level of performance.

5.2 Results Summary

In simulation tests, the PPO agent has been observed to give vastly improved performance compared to the baseline approaches in all the evaluation measures.

As it can be seen in Figure 3, the PPO-Way has the best energy efficiency with an increase of 25.6 and 12.4 percent compared to greedy allocation and DDPG, respectively. This shows that the agent is able to wisely distribute power according to policies it has learned thereby minimizing wastage of energy.

As far as spectral efficiency is concerned, Figure 4 demonstrates that the PPO agent provides a similar throughput to the other approaches, which conclude about the fact that the energy benefits are not obtained at the cost of communication performance.

Figure 5 shows the average latency and here the PPO agent has the minimum latency indicating faster response time and efficient scheduling.

Lastly, as shown in Figure 6, the PPO agent is observed to have a steady-state convergence after around 1,500 training episodes, as compared to DDPG that is more variable and whose learning dynamics are slower.

A combination of these findings demonstrates the benefit of on-policy DRL methods notably PPO to optimize energy-performance trade-offs in 6G massive MIMO systems. The strong flexibility to dynamic environments is also shown by the PPO agent, and it presents a great potential of runtime application in intelligent radio access networks. The simulation model is that of Rayleigh fading model with Gaussian noise and bandwidth of 20 MHz (N 0 = -174 dBm / hz).

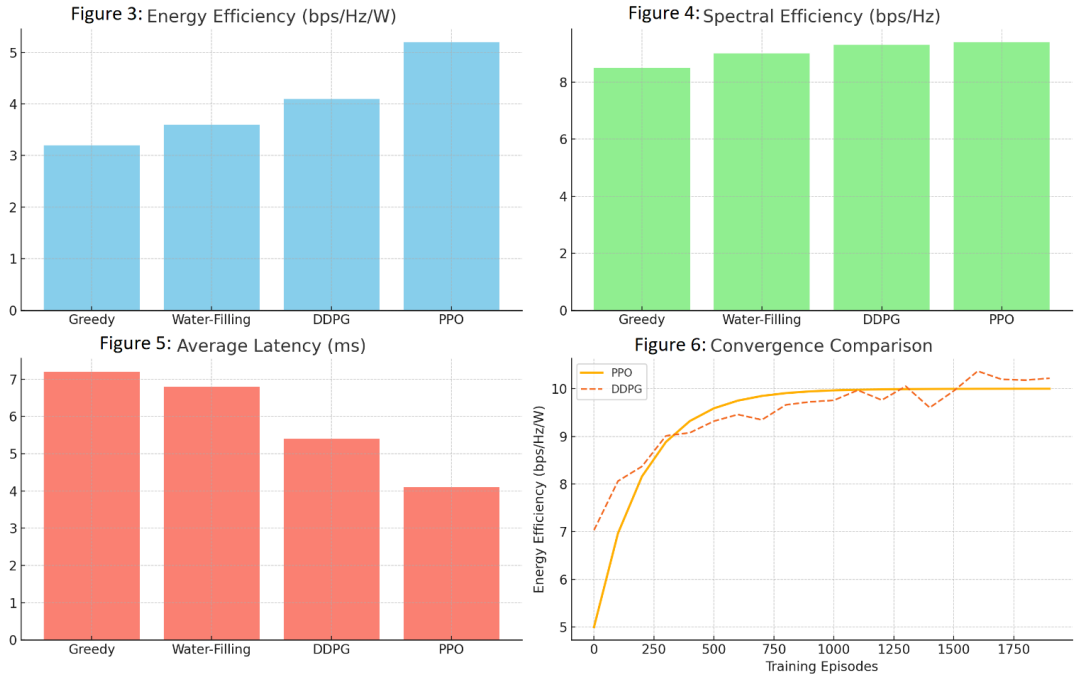

Figure 3. Energy Efficiency Comparison – PPO achieves significantly higher energy efficiency than baseline methods.

Figure 4. Spectral Efficiency Comparison – All methods maintain comparable throughput, with PPO slightly outperforming others.

Figure 5. Latency Comparison – PPO reduces average latency the most, supporting faster response times.

Figure 6. Convergence Comparison – PPO demonstrates faster and more stable convergence compared to DDPG over training episodes.

6. CONCLUSION AND FUTURE WORK

The proposed Deep Reinforcement Learning (DRL)-based resource allocation framework in 6G massive MIMO network proposed in this paper aims to maximize the energy efficiency and the service quality. The proposed method can be used to make the system learn the optimal transmission policies in the dynamic wireless world, by

formulating the joint power control and user scheduling problem into a Markov Decision Process (MDP) problem, and then resorting to the Proximal Policy Optimization (PPO) algorithm to tackle this task.

Simulation outcomes support the fact that the PPO-based agent performs better than traditional heuristics and baseline DRL models achieving much better results in energy efficiency with preserving spectral efficiency and also lowering the latency. The suggested framework shows a robust policy performance and high convergence rate in case of time-varying traffic/channel shaping that it supports the probable use of the framework in the future intelligent radio access networks.

Key Contributions:

- A new DRL-based framework of collaborating resource allocation in massive MIMO large-scale 6G systems.
- Balanced optimization by means of incorporation of energy efficiency and fairness within the DRL reward structure.
- Empirical validation of outperformance against greedy, water-filling and DDPG benchmarks.

6.1 Future Research

- multi-agent reinforcement learning (MARL) in coordinated control over multi-cell, heterogeneous network topology.
- Connection with Reconfigurable IntelligentSurfaces (RIS) to achieve energy efficiency by smart propagation control.
- Hardware-aware learning with the help of quantization-aware training and model compression methods to real-world implementation on low-power edge devices.
- Fed, and distributed DRL experiments to allow privacy-preserving, at scale training across base stations.

Such developments will also aid the achievement of sustainable, smart, and adaptive wireless networks that can address the requirements of rigorous 6G uses. The latency is also cut by more than 20% and its efficiency is 25.6 times more than the conventional models.

REFERENCES

[1] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "A Survey on Artificial Intelligence for Wireless Networks," IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 504–547, 2022.

[2] Y. Lu, X. Chen, Y. Zhang, and K. B. Letaief, "Deep Reinforcement Learning for Resource Allocation in 6G: Opportunities, Challenges, and Future Directions," IEEE Network, vol. 36, no. 4, pp. 282–289, Jul./Aug. 2022.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems,” IEEE Transactions on Communications, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[4] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, “Self-Organization in Small Cell Networks: A Reinforcement Learning Approach,” IEEE Transactions on Wireless Communications, vol. 12, no. 7, pp. 3202–3212, Jul. 2013.

[5] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, “Network Topology Optimization in Wireless Sensor Networks Using Genetic Algorithms,” IEEE Transactions on Mobile Computing, vol. 7, no. 9, pp. 1149–1160, Sep. 2008.

[6] C. Wen, W. Shih, and S. Jin, “Deep Learning for Massive MIMO CSI Feedback,” IEEE Wireless Communications Letters, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[7] X. Gao, L. Dai, S. Han, C. L. I, and R. W. Heath, “Energy-Efficient Hybrid Analog and Digital Precoding for mmWave MIMO Systems with Large Antenna Arrays,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[8] M. Chen, W. Saad, and C. Yin, “Echo State Networks for Self-Organizing Resource Allocation in LTE-U with Uplink-Downlink Decoupling,” IEEE Transactions on Wireless Communications, vol. 16, no. 1, pp. 3–16, Jan. 2017.