Information about the authors

Shakhmaran Seilov – Cand. Sc. (Eng.), Professor, Department of Computer and Software Engineering, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: seilov_shzh@enu.kz. ORCID: https://orcid.org/0000-0001-7057-0461.

Bibinur Zhursinbek – PhD student, Department of Computer and Software Engineering, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: zhursinbek99@gmail.com. ORCID: https://orcid.org/0009-0002-4505-470X.

Anatoliy Kremenchutskiy – Cand. Sc. (Eng.), Director of the India-Kazakhstan Centre of Excellence in ICT, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: ankrem@gmail.com. ORCID: https://orcid.org/0000-0003-4179-7880.

Saule Kazybaeva – Cand. Sc. (Agric.), Academician of the National Academy of Agrarian Sciences of Kazakhstan, LLP «Kazakh Research Institute of Fruit and Vegetable Growing», Almaty Region, Kazakhstan. ORCID: https://orcid.org/0000-0002-6833-0466.

Adil Maidanov* — Master, Lecturer, Department of Computer and Software Engineering, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: makeadil@mail.ru. ORCID: https://orcid.org/0000-0003-2392-5164.

Поступила в редакцию 02.06.2026
Поступила после доработки 09.09.2026
Принята к публикации 11.09.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-31

CC BY 4.0
Check for updates

ISTIR: 20.15.05

S. Aibagarov*, D. Marlambekov, A. Mukhanbet
Al-Farabi Kazakh National University,
71 Al-Farabi Ave., Almaty, Republic of Kazakhstan, 050040
*e-mail: awer1307dot@gmail.com

SAC-BASED UPLINK POWER CONTROL IN 5G NR: DEEP REINFORCEMENT LEARNING FOR JOINT THROUGHPUT AND ENERGY EFFICIENCY OPTIMIZATION

Abstract: *In this paper, the application of the Soft Actor-Critic (SAC) approach to uplink power control in 5G New Radio networks is investigated. A Gymnasium-compatible simulator for a 10-user uplink cell is constructed using the 3GPP TR 38.901 Urban Macro NLOS path-loss model with stochastic shadowing and Rayleigh fading at 3.5 GHz. The SAC agent observes SINR, transmit power, and path loss for all users and jointly selects continuous transmit powers in the range [0, 30] dBm. The training reward combines throughput (0.50), an energy-efficiency term (0.25), SINR quality (0.15), and fairness (0.10). Three SAC policies are trained for 500,000 timesteps and evaluated on 100 common scenarios against Full Power, Equal Power, and a simplified 3GPP-inspired Fractional Power Control (FPC) baseline. Using the standard network-level energy-efficiency metric, SAC achieves 14.866 Mbit/J at a mean transmit power of 9.905 dBm and a mean throughput of 2.540 Mbps. Network energy efficiency is significantly higher than Full Power and simplified FPC, while its difference from Equal Power is not statistically significant after multiplicity correction. A reward-weight sensitivity study confirms that changing objective priorities shifts the learned throughput-energy-efficiency operating point rather than yielding one universally dominant configuration.*

Key words: *5G NR, uplink power control, soft actor-critic, energy efficiency, SINR, Jain's fairness index, 3GPP channel model, deep reinforcement learning.*

1. Introduction

The quick development of connected devices in the framework of the 5G NR technology made the problem of radio resource management in the uplink much more complicated. In contrast to downlink scheduling when the gNB can centrally manage the process of transmission, the process of uplink power control implies that each User Equipment should make the decision about its transmit power using partial channel state information. This decentralized character of the problem, interference between users and time varying fading make it one of the most challenging open problem in the sphere of cellular network optimization [1, 2].

©S. Aibagarov, D. Marlambekov, A. Mukhanbet, 2026

On the other hand, there is an opposition between the two objectives of the uplink power control which are maximizing the sum rate and minimizing the total radiated power. The high level of the transmit power will help to increase the receive power of every UE but, at the same time, it will provide more interference for all the others and decrease their SINR. Due to the effect of interference coupling, the use of full transmit power will provide inefficient usage of energy and SINR balancing. The problem will be even more complicated due to the time varying fading channel [3, 4].

The solutions to the above problems offered by classical power control schemes involve using static or parameterized schemes. The scheme Full Power always works in such a way that it uses the maximum possible power P_max which guarantees maximization of throughput rate although results in too much interference. The Equal Power scheme uses an intermediate static value of power resulting in moderate throughput rate but without consideration of channel. The scheme 3GPP Fractional Power Control (FPC) [5] partially solves the problem of path losses but does not consider fast Rayleigh fading and interference [5, 6].

Deep reinforcement learning (DRL) has become a promising framework for adaptive management of resources in wireless communication networks [7, 17]. SAC is one of the DRL algorithms that has proven to be especially effective at solving tasks of a continuous nature, such as power allocation, thanks to its off-policy nature, which allows learning from a replay memory, and its maximum-entropy objective that drives exploration and avoids converging into local minima prematurely [18], [9]. Learning-based techniques have been shown to work effectively in downlink beamforming and spectrum management tasks [10, 11, 19]; however, joint optimization of the parameters mentioned is not yet widely investigated.

There are multiple common issues in existing DRL-based works on power control algorithms. Firstly, most of the previous work uses simplistic channel models without any real-world path loss and, thus, significantly underestimates the heterogeneity between different UEs [12]. Secondly, most of the literature focuses on optimizing just one metric, while in reality, operators have to deal with multiple metrics that may conflict with each other [13, 20]. Finally, the comparison with standard 3GPP benchmarks like FPC is not used in most cases and, therefore, no practical benefit from using DRL can be evaluated [6].

Recent studies further demonstrate the growing use of DRL for power- and resource-control problems. SAC has been applied to joint power and resource allocation in O-RAN-enabled cell-free massive MIMO systems [21], while SAC-based adaptive energy management has also been studied for NB-IoT power-saving control [22]. Other recent work investigates DRL-based power allocation for energy- and spectral-efficiency improvement in dense 5G deployments [23]. These studies confirm the relevance of adaptive learning-based control, but they address different architectures and objectives. The present work is intentionally narrower: it evaluates one continuous-control SAC formulation for joint uplink transmit-power selection in a controlled 3GPP-based simulator against conventional reference strategies rather than providing a comprehensive benchmark of DRL algorithms.

Luong et al. [7] and Ye et al. [14] discuss those issues and point out the need for further research in those areas in their surveys. The main question addressed in this paper is the following: can SAC agent learn a power control algorithm that will be better than Full Power, Equal Power, and FPC benchmarks in energy efficiency and SINR fairness, while still having competitive aggregate throughput, in realistic 3GPP UMa NLOS channel environment?

The scientific novelty of this study lies in combining centralized continuous multi-user power control with a four-objective reward and a reproducible 3GPP-based stochastic channel model, while directly comparing the learned policy with conventional reference power-control strategies under common evaluation scenarios. Unlike studies focused on broader O-RAN slicing, power-saving modes, or multi-algorithm benchmarking, the proposed setup isolates the throughput-energy-efficiency trade-off of joint uplink transmit-power adaptation for ten heterogeneous UEs. The resulting conclusions are therefore restricted to the considered simulator and baselines rather than generalized to all DRL or static power-control methods.

Contributions of this work are as follows. Firstly, we construct a Gymnasium-compatible, simplified interference-coupled uplink power-control scenario based on 3GPP TR 38.901 UMa NLOS path loss at 3.5 GHz, with ten UEs and a 9 MHz effective system bandwidth. Secondly, we design a combined SAC reward with four terms: throughput (0.50), energy efficiency (0.25), SINR quality (0.15), and fairness (0.10), and we additionally evaluate the sensitivity of the learned policy to three

alternative weight settings. Thirdly, three independently trained SAC policies are evaluated on 100 common scenarios against Full Power, Equal Power, and a simplified 3GPP-inspired FPC baseline using standard network energy efficiency and Jain's Fairness Index computed from linear SINR. Finally, we provide paired non-parametric statistical comparisons and seed-level robustness analysis to distinguish mean differences from statistically supported effects.

The rest of the paper is organized as follows. Section 2 introduces system model, MDP and training process. Section 3 discusses experimental results. Section 4 summarizes with suggestions for future work.

2. Methods and Materials

2.1 System Model

We consider a single-cell 5G NR-inspired uplink scenario with one gNB and N = 10 UEs over a total effective bandwidth B = 9 MHz, corresponding to 50 PRBs at 180 kHz each. The simulator does not explicitly schedule orthogonal PRBs among UEs. Instead, it uses a simplified interference-coupled shared-spectrum abstraction: all UEs transmit simultaneously and contribute to the interference term, while B/N is used as the effective per-user bandwidth in the Shannon-rate calculation. Thus, the model is intended as a controlled power-control abstraction rather than a complete NR uplink scheduler. Each UE independently selects its transmit power in the range [0, 30] dBm at every time step. The received SINR at the gNB for UE i is:

$$SINR_i = \frac{P_i G_i}{N_0 + \sum_{j \neq i} P_j G_j} \quad (1)$$

where G_i is the composite channel gain of UE i and N_0 is the thermal noise power. The noise floor is computed as:

$$N_0 = k_B \cdot T \cdot B \cdot NF \quad (2)$$

where k is Boltzmann's constant, T = 290 K, and NF = 9 dB is the receiver noise figure, giving approximately −95.46 dBm over the 9 MHz band. The instantaneous throughput of UE i is modeled using the Shannon capacity formula applied to its effective equal bandwidth share:

$$R_i = \frac{B}{N} \cdot log_2(1 + SINR_i) \ [bps] \quad (3)$$

During training, the energy-efficiency reward term is defined as the aggregate per-user quantity obtained by summing the individual ratios:

$$EE_{agg} = \sum_{i=1}^{N} \frac{R_i}{P_i} \quad (4)$$

where transmit power is converted from dBm to Watts. For performance evaluation, energy efficiency is measured at the network level as:

$$EE_{network} = \frac{\sum_i R_i}{\sum_i P_i} \quad (5)$$

with:

$$P_i[W] = 10^{\frac{P_i[dBm]-30}{10}} \quad (6)$$

The network-level metric is used for all system-level energy-efficiency comparisons reported in this study.

$$EE_i = \frac{R_i}{P_i} \ [bps/W] \quad (7)$$

2.2. Channel Model

The channel is modeled according to the 3GPP TR 38.901 Urban Macro (UMa) NLOS model at f=3.5 GHz [12]. The path loss for user equipment i at distance d_i ∈ [50, 500] is:

$$PL_i = 13.54 + 39.08 \cdot log_{10}(d_i) + 20 \cdot log_{10}(f) - 0.6(h_{UT} - 1.5)\ [dB] \quad (8)$$

where h_{BS} = 25 m and h_{UT} = 1.5 m. The composite linear channel gain per TTI combines three components:

$$G_i = 10^{\frac{-(PL_i + \xi_i)}{10}} \cdot |h_i|^2 \quad (9)$$

where the log-normal shadowing component is distributed as N(0, 8²) dB and the fast-fading component follows a Rayleigh model. Both stochastic channel components are redrawn at each environment step, whereas UE distances are sampled at episode initialization and remain fixed for the 200-step episode. This produces persistent large-scale distance heterogeneity with step-wise stochastic channel variation.

2.3. State Space, Action Space, and Reward

The observation vector s_t is a 30-dimensional normalized feature vector comprising three 10-dimensional blocks per UE:

$$s_t = [SINR_{norm}(10),\ P_{norm}(10),\ PL_{norm}(10)] \in [0,1]^{30} \quad (10)$$

SINR normalization uses the range [−20, 40] dB; power normalization divides by P_{max} = 30 dBm; path-loss normalization spans the range corresponding to d ∈ [50, 500] m. Including path loss allows the agent to distinguish between near and far UEs. The action space A = $[-1,1]^{10}$ is a continuous vector mapped to transmit powers as:

$$P_i = \frac{a_i + 1}{2} \cdot P_{max}\ [dBm] \quad (11)$$

so that a_i = −1 corresponds to 0 dBm and a_i = +1 to 30 dBm. Continuous actions allow fine-grained power adjustments impossible with discrete action spaces [19]. The composite reward at each TTI is:

$$r_t = 0.50 \cdot \tilde{r}_{tp} + 0.25 \cdot \tilde{r}_{ee} + 0.15 \cdot \tilde{r}_{sinr} + 0.10 \cdot Ÿ(SINR) \quad (12)$$

where the normalized throughput and energy efficiency components are:

$$\tilde{r}_{tp} = \frac{\sum R_i}{R_{max}}, \quad \tilde{r}_{ee} = \frac{\sum EE_i}{EE_{max}} \quad (13)$$

and the SINR quality component clips mean SINR (dB) to the practical range [−35, −15] dB:

$$\tilde{r}_{sinr} = clip\left(\frac{\overline{SINR} - (-35)}{-15 - (-35)}, 0, 1\right) \quad (14)$$

with normalization ceilings R_{max} = 5 Mbps and EE_{max} = 500 Mbps/J. The fairness term uses Jain's Fairness Index [15] over per-UE SINR values:

$$Ÿ(SINR) = \frac{(\sum SINR_i)^2}{N \cdot \sum SINR_i^2} \in \left[\frac{1}{N}, 1\right] \quad (15)$$

A value of J = 1 corresponds to perfectly equal SINR distribution across UEs. The SINR component in Equation (9) penalizes excessive power reduction that degrades link quality below baseline levels, preventing degenerate low-power solutions.

The reward weights were selected heuristically to reflect the relative operational importance of the four objectives while keeping their normalized contributions comparable. Throughput receives the largest weight (0.50) because maintaining user data rate is the primary QoS objective; energy

efficiency is assigned 0.25 to strongly discourage unnecessarily high transmit power. The SINR term (0.15) acts as a link-quality safeguard against overly aggressive power reduction, while the fairness term (0.10) prevents the policy from improving aggregate performance by disproportionately degrading individual UEs. This ordering therefore prioritizes service performance and energy savings while retaining minimum link quality and inter-user balance.

For the evaluation metrics, Jain's Fairness Index is computed using positive linear SINR values. SINR values expressed in dB are first converted to the linear scale as:

$$SINR_{linear} = 10^{\frac{SINR_{dB}}{10}} \quad (16)$$

after which Jain's Fairness Index is calculated as:

$$J = \frac{(\sum_i x_i)^2}{N \sum_i x_i^2} \quad (17)$$

where x_i denotes the linear SINR of UE i. The same linear representation of SINR is used in the fairness component of the training reward.

2.4. SAC Agent and Training

The SAC agent [9] is implemented using Stable-Baselines3 [16] with an MlpPolicy comprising two hidden layers of 128 units (Table 1). SAC's maximum entropy objective augments the reward with the policy entropy, encouraging exploration and preventing mode collapse:

$$\pi^* = arg\,max\pi \sum {}^{\circ}c_{\pi}[r_t + \alpha \cdot \mathcal{H}(\pi(\cdot \mid s_t))] \quad (18)$$

where H denotes entropy and α is a temperature parameter tuned automatically. Key hyperparameters are listed in Table 1.

Table 1 – SAC Hyperparameters.

Hyperparameter	Value
Algorithm	SAC (off-policy, continuous)
Total timesteps	500,000 per seed
Parallel environments	4
Network architecture	[128, 128] MLP (actor & critic)
Learning rate	3×10^{-4}
Replay buffer size	300,000
Mini-batch size	4,096
Soft update (τ)	0.005
Entropy coefficient	auto (target entropy)
Learning starts	5,000
Random seeds	0, 1, 2

In training, we use N_envs = 4 parallel environments in order to speed up data acquisition. The experience buffer holds the latest 300,000 transitions, and the gradients are updated after every step. Another evaluation environment (seed + 100) checks for the performance of the policy after every 25,000 steps.

The experiments were implemented in Python 3.11 using Gymnasium 0.29.1, Stable-Baselines3 2.3.2, and PyTorch 2.6.0. Training was performed on a CUDA-enabled workstation equipped with an NVIDIA GeForce RTX 5080 GPU; the implementation automatically falls back to CPU execution if CUDA is unavailable. The software dependencies and training scripts were fixed across all random seeds to ensure reproducibility.

2.5. Baseline Algorithms

Three conventional reference approaches to power control are considered as benchmarks. Full Power: all UEs transmit at the maximum power Pmax = 30 dBm irrespective of channel condition. Equal Power: all UEs transmit at a constant P = 15 dBm. Simplified 3GPP-inspired Fractional Power Control (FPC): the implemented baseline retains the principal fractional path-loss compensation term but is not the complete TS 38.213 PUSCH power-control procedure. It omits

explicit RB-allocation/numerology scaling, transport-format correction, TPC/closed-loop state, and BWP/reference-signal configuration terms. The simplified baseline sets the transmission power equal to:

$$P_i = clip(P_0 + \alpha \cdot PL_i, 0, 30) \ [dBm] \quad (19)$$

with open-loop parameters P_0 = −60 dBm and α = 0.8. This simplified FPC baseline compensates for deterministic path-loss differences but does not implement the complete standardized 3GPP uplink power-control chain or react to the instantaneous stochastic channel realization.

2.6. Evaluation Protocol

Each baseline is evaluated over 100 common scenarios of 200 TTIs. Each of the three independently trained SAC policies (training seeds 0, 1, and 2) is evaluated on the same 100 scenario seeds; the primary SAC result is formed by averaging the three policy outcomes within each scenario, yielding 100 paired scenario-level observations while retaining separate seed-level summaries. Learning-curve checkpoints use 10 evaluation episodes, whereas Figure 4 uses 50 episodes of the seed-0 SAC policy only to visualize per-UE power distributions. Performance is reported using mean and sample standard deviation for descriptive results, with median and IQR for the strongly skewed network-energy-efficiency metric. Inferential comparisons use two-sided paired Wilcoxon signed-rank tests on common scenarios, Holm correction across nine tests, and 95% bootstrap confidence intervals where reported. The evaluation metrics include mean SINR, aggregate throughput, standard network energy efficiency, mean transmit power, and Jain's Fairness Index computed from linear SINR.

2.7. Reward-Weight Sensitivity Analysis

To assess whether the conclusions depend critically on the manually selected reward weights, a limited sensitivity study evaluates four configurations using the same SAC architecture and training length. Configuration A is the original setting (0.50/0.25/0.15/0.10 for throughput/energy efficiency/SINR/fairness). Configuration B is throughput-oriented (0.60/0.20/0.10/0.10), Configuration C is energy-efficiency-oriented (0.40/0.35/0.15/0.10), and Configuration D is more balanced toward energy efficiency and fairness (0.40/0.30/0.15/0.15). Configurations B-D are each trained with three random seeds for 500,000 timesteps and all twelve policies are evaluated on the same 100 scenario seeds. The purpose is sensitivity analysis of the operating point, not proof that any weight vector is globally optimal.

3. Results

3.1. Learning Dynamics

The SAC learning curve is shown in Figure 1 for 500,000 training timesteps. The plotted quantity is the accumulated return over a complete 200-step evaluation episode, averaged across the three training seeds; the shaded region represents ±1 standard deviation across seeds. At 500,000 timesteps the three episode returns are 89.85, 82.58, and 80.06, with a mean of 84.16, corresponding to approximately 0.421 reward per step. Over the 200,000-350,000 timestep interval, the mean episode return is approximately 82.62 (0.413 per step), showing that the policy has reached a broadly stable training regime.



Figure 1 – SAC learning curve averaged over three training seeds. The y-axis is the mean accumulated return over a 200-step evaluation episode; the shaded area denotes ±1 standard deviation across seeds

3.2. Comparative Performance

Table 2 summarizes the performance comparison using the standard network-level energy-efficiency metric and Jain's Fairness Index computed from linear SINR. SAC achieves a mean throughput of 2.540 Mbps, a network energy efficiency of 14.866 Mbit/J, and a mean transmit power of 9.905 dBm. The corresponding network-energy-efficiency means are 0.268 Mbit/J for Full Power, 7.486 Mbit/J for Equal Power, and 0.261 Mbit/J for simplified FPC. SAC achieves substantially higher mean network energy efficiency than Full Power and simplified FPC. Its mean network energy efficiency is also higher than that of Equal Power; however, this difference is not statistically significant after Holm correction. Because the SAC distribution is strongly right-skewed, both median/IQR and inferential statistics are considered alongside the mean values.

Table 2 – Performance comparison over 100 paired evaluation scenarios. Throughput and network EE are reported as mean ± SD; SINR, mean power, and Jain fairness are scenario means.

Method	SINR (dB)	TP (Mbps)	Network EE (Mbit/J)	Power (dBm)	Jain's FI (linear SINR)
SAC (ours)	−26.971	2.540 ± 1.169	14.866 ± 25.347	9.905	0.166
Full Power	−23.356	2.682 ± 0.556	0.268 ± 0.056	30.000	0.163
Equal Power	−23.864	2.367 ± 0.543	7.486 ± 1.717	15.000	0.167
Simplified FPC	−19.553	2.098 ± 0.091	0.261 ± 0.034	28.094	0.198

Simplified FPC attains the highest mean SINR (−19.553 dB) because its path-loss compensation drives many users toward high transmit powers. SAC operates at a lower mean SINR (−26.971 dB) while reducing mean transmit power to 9.905 dBm. This reduction in absolute SINR is therefore an expected consequence of the learned low-power operating point rather than, by itself, evidence of ineffective control.

The trade-off is acceptable in the considered simulator because SAC preserves a mean throughput of 2.540 Mbps while achieving substantially higher standard network energy efficiency than Full Power and simplified FPC. Relative to Equal Power, SAC has a higher mean network EE, but the paired difference is not statistically significant after Holm correction; this comparison is therefore reported as a numerical operating-point difference rather than a proven superiority claim.

Using Jain's Fairness Index computed from linear SINR, the mean fairness values are 0.166 for SAC, 0.163 for Full Power, 0.167 for Equal Power, and 0.198 for simplified FPC. Thus, SAC does not outperform all baselines in terms of fairness. Paired tests show no significant fairness difference between SAC and Full Power or Equal Power, while simplified FPC achieves significantly higher fairness.

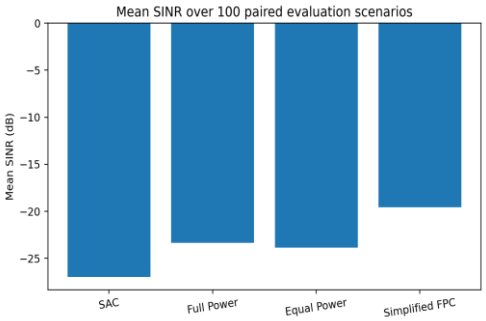


Figure 2 – Mean SINR by method over 100 paired evaluation scenarios. Detailed uncertainty estimates and paired statistical comparisons are reported in the accompanying analysis

3.3. Throughput-Energy Efficiency Tradeoff

Figure 3 plots mean throughput against the standard network-level energy-efficiency metric. Full Power provides the highest observed mean throughput (2.682 Mbps) but only 0.268 Mbit/J network EE. Equal Power yields 2.367 Mbps and 7.486 Mbit/J. Simplified FPC yields 2.098 Mbps and 0.261 Mbit/J. SAC occupies an intermediate throughput point at 2.540 Mbps while reaching 14.866 Mbit/J mean network EE. This operating point is not described as globally Pareto-optimal or unreachable by other policies; it is simply not achieved by the three considered baseline methods.

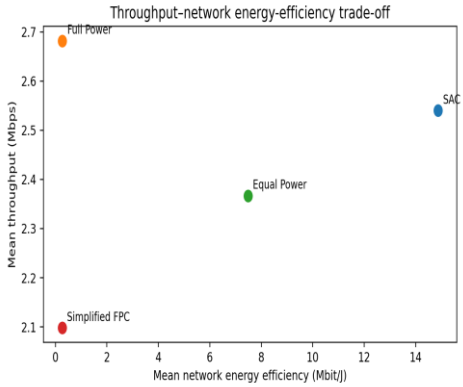


Figure 3 – Throughput versus standard network energy efficiency. Each point is the mean over 100 paired evaluation scenarios; the figure summarizes the operating points of the four considered methods

3.4. Per-UE Power Allocation

Per-UE transmit powers allocated by the SAC agent are shown in Figure 4 using 50 episodes of the seed-0 policy for visualization. The learned allocation is heterogeneous across UEs, reflecting the centralized policy's joint response to path-loss and previously observed SINR information. Because the stochastic channel is redrawn at each environment step, the policy should be interpreted as state-adaptive under the implemented observation/update timing rather than as a complete instantaneous NR scheduler.

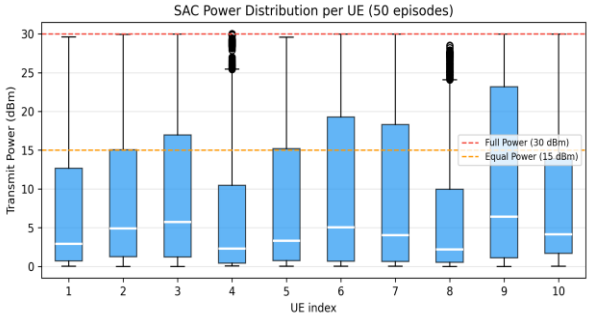


Figure 4 – Per-UE transmit-power distribution for the seed-0 SAC policy over 50 visualization episodes of 200 steps each

3.5. Reward-Weight Sensitivity

The limited sensitivity experiment confirms that the reward weights materially change the learned operating point. Configuration A (0.50/0.25/0.15/0.10) yields 2.540 Mbps throughput and 14.866 Mbit/J network EE. Increasing the throughput weight in Configuration B raises throughput to 4.836 Mbps but reduces network EE to 5.277 Mbit/J and fairness to 0.120. The energy-efficiency-oriented Configuration C reaches 19.859 Mbit/J with 1.570 Mbps throughput, while the more balanced Configuration D reaches 18.688 Mbit/J with 1.593 Mbps and fairness 0.190. No alternative dominates Configuration A across all metrics. The experiment therefore supports the interpretation that the manually selected weights encode an explicit trade-off rather than a globally optimal setting (Table 3, Figure 5).

Table 3 – Reward-weight sensitivity over 100 common evaluation scenarios.

Configuration	Weights TP/EE/SINR/Fair	TP (Mbps)	Network EE (Mbit/J)	Fairness
A Original	.50/.25/.15/.10	2.540 ± 1.169	14.866 ± 25.347	0.166
B TP-oriented	.60/.20/.10/.10	4.836 ± 1.486	5.277 ± 2.502	0.120
C EE-oriented	.40/.35/.15/.10	1.570 ± 0.298	19.859 ± 19.538	0.189
D More balanced	.40/.30/.15/.15	1.593 ± 0.332	18.688 ± 20.831	0.190

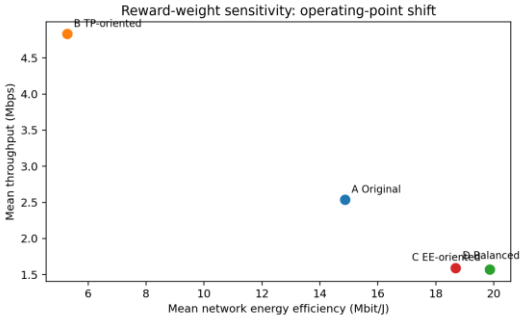


Figure 5 – Reward-weight sensitivity. Changing the reward priorities shifts the learned throughput–network-energy-efficiency operating point; no configuration dominates across all reported metrics

4. Conclusion

In this work, we investigated Soft Actor-Critic for joint uplink transmit-power control in a simplified interference-coupled 5G NR-inspired simulator using 3GPP TR 38.901 UMa NLOS path loss and stochastic channel variation. The centralized SAC policy jointly selects ten continuous transmit powers from [0, 30] dBm using a multi-objective reward that balances throughput, energy efficiency, SINR quality, and fairness.

Across 100 common evaluation scenarios per policy, SAC achieves 2.540 Mbps mean throughput, 14.866 Mbit/J standard network energy efficiency, and 9.905 dBm mean transmit power. Network EE is significantly higher than Full Power and simplified FPC after Holm correction, while its difference from Equal Power is not statistically significant. Jain fairness is not superior to all baselines; simplified FPC achieves the highest mean fairness among the considered methods. The reward-weight sensitivity experiment further shows that throughput, energy efficiency, SINR, and fairness can be shifted substantially by changing objective priorities, with no tested configuration dominating the others across all metrics.

From a practical perspective, the results indicate that continuous DRL control can provide low-power operating points that preserve competitive throughput while substantially improving network energy efficiency relative to high-power reference strategies in the considered simulator. Such behavior is relevant to adaptive uplink resource-management systems in which transmit-power decisions must balance service performance and energy consumption. The present study should be viewed as evidence for this design direction rather than as proof of superiority over all conventional or learning-based power-control algorithms.

Several limitations define the scope of these conclusions. First, the simulator uses a shared-spectrum interference abstraction and does not explicitly schedule orthogonal NR PRBs or model inter-cell interference. Second, the implemented FPC is a simplified 3GPP-inspired path-loss-compensation baseline rather than the complete TS 38.213 procedure. Third, shadowing and Rayleigh components are redrawn at each environment step, and the centralized SAC policy was trained and evaluated under the same model family. Fourth, the study intentionally compares SAC with conventional reference strategies rather than TD3, PPO, DDPG, or other DRL controllers; a carefully tuned multi-algorithm benchmark is left for future work. Finally, the reward-weight ablation demonstrates sensitivity rather than global optimality. Future research should incorporate explicit NR PRB scheduling, inter-cell interference, more realistic temporally correlated channels, complete standardized baselines, controlled DRL comparisons, and simulation-to-reality validation.

References

1. Shen Z. Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints / Z. Shen, J.G. Andrews, B.L. Evans // IEEE Transactions on Wireless Communications. – 2005. – Vol. 4, № 6. – P. 2726-2737. https://doi.org/10.1109/TWC.2005.858010.

2. Tse D. Fundamentals of Wireless Communication / D. Tse, P. Viswanath. – Cambridge: Cambridge University Press, 2005. – 564 p. https://doi.org/10.1017/CBO9780511807213.

3. Tombaz S. Energy- and cost-efficient ultra-high-capacity wireless access / S. Tombaz, A. Västberg, J. Zander // IEEE Wireless Communications. – 2011. – Vol. 18, № 5. – P. 18-24. https://doi.org/10.1109/MWC.2011.6056688.

4. Zander J. Performance of optimum transmitter power control in cellular radio systems / J. Zander // IEEE Transactions on Vehicular Technology. – 1992. – Vol. 41, № 1. – P. 57-62. https://doi.org/10.1109/25.120145.
5. 3GPP TS 38.213. NR; Physical layer procedures for control. Version 17.4.0 [Electronic resource]. – 3rd Generation Partnership Project, 2022. https://www.3gpp.org/dynareport/38213.htm.
6. Holtkamp H. Minimizing base station power consumption / H. Holtkamp, G. Auer, S. Bazzi, H. Haas // IEEE Journal on Selected Areas in Communications. – 2014. – Vol. 32, № 2. – P. 297-306. https://doi.org/10.1109/JSAC.2013.131009.
7. Applications of deep reinforcement learning in communications and networking: A survey / N.C. Luong et al // IEEE Communications Surveys & Tutorials. – 2019. – Vol. 21, № 4. – P. 3133-3174. https://doi.org/10.1109/COMST.2019.2916583.
8. Application of machine learning in wireless networks: Key techniques and open issues / Y. Sun et al // IEEE Communications Surveys & Tutorials. – 2019. – Vol. 21, № 4. – P. 3072-3108. https://doi.org/10.1109/COMST.2019.2924243.
9. Haarnoja T. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor / T. Haarnoja, A. Zhou, P. Abbeel, S. Levine // Proceedings of the 35th International Conference on Machine Learning (ICML 2018). – 2018. – Vol. 80. – P. 1861-1870. https://proceedings.mlr.press/v80/haarnoja18b.html.
10. Cui W. Spatial deep learning for wireless scheduling / W. Cui, K. Shen, W. Yu // IEEE Journal on Selected Areas in Communications. – 2019. – Vol. 37, № 6. – P. 1248-1261. https://doi.org/10.1109/JSAC.2019.2904352.
11. Deep reinforcement learning for radio resource allocation and management in next generation heterogeneous wireless networks: A survey / A. Alwarafy et al // IEEE Open Journal of the Communications Society. – 2022. – Vol. 3. – P. 1967-1986. https://doi.org/10.1109/OJCOMS.2022.3153226.
12. 3GPP TR 38.901. Study on channel model for frequencies from 0.5 to 100 GHz. Version 17.0.0 [Electronic resource]. – 3rd Generation Partnership Project, 2022. https://www.3gpp.org/dynareport/38901.htm.
13. Naparstek O. Deep multi-user reinforcement learning for distributed dynamic spectrum access / O. Naparstek, K. Cohen // IEEE Transactions on Wireless Communications. – 2019. – Vol. 18, № 1. – P. 310-323. https://doi.org/10.1109/TWC.2018.2878578.
14. Ye H. Deep reinforcement learning based resource allocation for V2V communications / H. Ye, G.Y. Li, B.-H.F. Juang // IEEE Transactions on Vehicular Technology. – 2019. – Vol. 68, № 4. – P. 3163-3173. https://doi.org/10.1109/TVT.2019.2897134.
15. Jain R. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems / R. Jain, D.-M. Chiu, W.R. Hawe. – Hudson: Digital Equipment Corporation, 1984. – 38 p. – (DEC Research Report TR-301). https://arxiv.org/abs/cs/9809099.
16. Stable-Baselines3: Reliable reinforcement learning implementations / A. Raffin et al // Journal of Machine Learning Research. – 2021. – Vol. 22, № 268. – P. 1-8. https://jmlr.org/papers/v22/20-1364.html.
17. Sutton R.S. Reinforcement Learning: An Introduction / R.S. Sutton, A.G. Barto. – 2nd ed. – Cambridge, MA : MIT Press, 2018. – 552 p. http://incompleteideas.net/book/the-book-2nd.html.
18. Human-level control through deep reinforcement learning / V. Mnih et al // Nature. – 2015. – Vol. 518, № 7540. – P. 529-533. https://doi.org/10.1038/nature14236.
19. Proximal policy optimization algorithms / J. Schulman et al // arXiv. – 2017. – arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347.
20. Machine learning paradigms for next-generation wireless networks / C. Jiang et al // IEEE Wireless Communications. – 2017. – Vol. 24, № 2. – P. 98-105. https://doi.org/10.1109/MWC.2016.1500356WC.
21. Eskandari M. Network slicing in O-RAN-enabled cell-free massive MIMO: A DRL-based power control / M. Eskandari, M. Rahmani Ghourtani, A.G. Burr // Proceedings of the 2025 IEEE Wireless Communications and Networking Conference (WCNC). – Milan: IEEE, 2025. https://doi.org/10.1109/WCNC61545.2025.10978801.
22. Anbazhagan S. Adaptive power-saving mode control in NB-IoT networks using soft actor-critic reinforcement learning for optimal power management / S. Anbazhagan, R.K. Mugelan // Scientific Reports. – 2025. – Vol. 15, Art. 34618. https://doi.org/10.1038/s41598-025-18214-4.

23. Base station power control strategy in ultra-dense networks via deep reinforcement learning / Q. Chen et al // Physical Communication. – 2025. – Vol. 71, Art. 102655. https://doi.org/10.1016/j.phycom.2025.102655.

Acknowledgment

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant № BR24993211).

С. Айбагаров*, Д. Марламбеков, А. Муханбет
Казахский национальный университет имени аль-Фараби,
Республика Казахстан, г. Алматы, 050040, проспект Аль-Фараби, 71
*e-mail: awer1307dot@gmail.com

УПРАВЛЕНИЕ МОЩНОСТЬЮ ВОСХОДЯЩЕГО КАНАЛА НА ОСНОВЕ SAC В 5G NR: ГЛУБОКОЕ ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ ДЛЯ СОВМЕСТНОЙ ОПТИМИЗАЦИИ ПРОПУСКНОЙ СПОСОБНОСТИ И ЭНЕРГОЭФФЕКТИВНОСТИ

В статье исследуется применение Soft Actor-Critic (SAC) для управления мощностью восходящего канала в сетях 5G NR. Разработан совместимый с Gymnasium симулятор для 10 пользователей на основе модели 3GPP TR 38.901 UMa NLOS. Агент SAC совместно определяет непрерывные мощности передачи в диапазоне [0, 30] дБм. Три независимо обученные политики SAC оценивались на 100 общих сценариях и сравнивались с Full Power, Equal Power и упрощенным 3GPP-inspired Fractional Power Control. При использовании стандартной сетевой метрики энергоэффективности SAC достигает 14,866 Мбит/с/Дж при средней мощности 9,905 дБм и пропускной способности 2,540 Мбит/с. Энергоэффективность статистически выше, чем у Full Power и упрощенного FPC, однако отличие от Equal Power после поправки на множественные сравнения статистически незначимо. Анализ чувствительности весов функции вознаграждения показывает, что изменение приоритетов смещает компромисс между пропускной способностью и энергоэффективностью и не выявляет единственной доминирующей конфигурации.

Ключевые слова: *5G NR, управление мощностью восходящего канала, алгоритм «мягкий актер-критик», энергоэффективность, SINR, индекс справедливости Джейна, модель канала 3GPP, глубокое обучение с подкреплением.*

S. Aibagarov*, D. Marlambekov, A. Mukhanbet
Әл-Фараби атындағы Қазақ ұлттық университеті,
Қазақстан Республикасы, Алматы, 050040, Әл-Фараби даңғылы, 71
*e-mail: awer1307dot@gmail.com

SAC НЕГІЗІНДЕГІ 5G NR ЖЕЛІЛЕРІНДЕГІ КӨТЕРІЛМЕЛІ (UPLINK) ҚУАТТЫ БАСҚАРУ: ӨТКІЗУ ҚАБІЛЕТІ МЕН ЭНЕРГИЯ ТИІМДІЛІГІН БІРЛЕСІП ОҢТАЙЛАНДЫРУҒА АРНАЛҒАН ТЕРЕҢ КҮШЕЙТІЛГЕН ОҚЫТУ

Бұл мақалада 5G NR желілеріндегі uplink тарату қуатын басқару үшін Soft Actor-Critic (SAC) әдісі зерттеледі. 3GPP TR 38.901 UMa NLOS моделіне негізделген 10 пайдаланушыға арналған Gymnasium-үйлесімді симулятор жасалды. SAC агенті [0, 30] дБм аралығындағы он пайдаланушының тарату қуатын бірлесіп басқарады. Үш тәуелсіз SAC саясаты 100 ортақ сценарийде бағаланып, Full Power, Equal Power және жеңілдетілген 3GPP-inspired Fractional Power Control әдістерімен салыстырылды. Стандартты желілік энергия тиімділігі метрикасы бойынша SAC 14,866 Мбит/с/Дж нәтижесіне, 9,905 дБм орташа қуатқа және 2,540 Мбит/с өткізу қабілетіне қол жеткізді. SAC энергия тиімділігі Full Power және жеңілдетілген FPC әдістерінен статистикалық тұрғыдан жоғары, ал Equal Power әдісімен айырмашылық көптік салыстыру түзетуінен кейін статистикалық мәнді емес. Марапат салмақтарының сезімталдық талдауы мақсат басымдықтары өзгерген кезде өткізу қабілеті мен энергия тиімділігі арасындағы жұмыс нүктесі өзгеретінін және барлық метрикалар бойынша бір ғана үстем конфигурация жоқ екенін көрсетеді.

Түйін сөздер: *5G NR, көтерілмелі қуатты басқару, soft actor-critic, энергия тиімділігі, SINR, Джейннің әділдік индексі, 3GPP арна моделі, терең күшейтілген оқыту.*

Information about the authors

Serik Aibagarov* – Bachelor student, Junior Reseacher, Computer Science department Almaty, Kazakhstan; e-mail: awer1307dot@gmail.com. ORCID: https://orcid.org/0009-0009-4946-4926.

Duman Marlambekov – PhD student, Junior Reseacher, Computer Science department, Almaty, Kazakhstan; e-mail: marlambekov_duman@live.kaznu.kz. ORCID: https://orcid.org/0009-0005-3266-2126.

Aksultan Mukhanbet – PhD student, Researcher, Computer Science department, Almaty, Kazakhstan; e-mail: mukhanbetaksultan0414@gmail.com. ORCID: https://orcid.org/0000-0003-4699-0436.

Информация об авторах

Серик Айбагаров* – студент, младший научный сотрудник, кафедра компьютерных наук Алматы, Казахстан; e-mail: awer1307dot@gmail.com. ORCID: https://orcid.org/0009-0009-4946-4926.

Думан Марламбеков – докторант, младший научный сотрудник, кафедра компьютерных наук, Алматы, Казахстан; e-mail: marlambekov_duman@live.kaznu.kz. ORCID: https://orcid.org/0009-0005-3266-2126.

Аксултан Муханбет – Студент докторантуры, научный сотрудник, кафедра компьютерных наук, Алматы, Казахстан; e-mail: mukhanbetaksultan0414@gmail.com. ORCID: https://orcid.org/0000-0003-4699-0436.

Авторлар туралы ақпарат

Серик Айбагаров* – Бакалавр студенті, кіші ғылыми қызметкер, Компьютерлік ғылымдар кафедрасы, Алматы, Қазақстан; e-mail: awer1307dot@gmail.com. ORCID: https://orcid.org/0009-0009-4946-4926.

Думан Марламбеков – Докторантура студенті, кіші ғылыми қызметкер, Компьютерлік ғылымдар кафедрасы, Алматы, Қазақстан; e-mail: marlambekov_duman@live.kaznu.kz. ORCID: https://orcid.org/0009-0005-3266-2126.

Ақсұлтан Муханбет – Докторантура студенті, ғылыми қызметкер, Компьютерлік ғылымдар кафедрасы, Алматы, Қазақстан; e-mail: mukhanbetaksultan0414@gmail.com. ORCID: https://orcid.org/0000-0003-4699-0436.

Received 09.06.2026
Revised 07.09.2026
Accepted 18.09.2026