Optimisation of Resource Allocation in Heterogeneous Wireless Networks Using Deep Reinforcement Learning

Oluwaseyi Giwa*, Jonathan Shock§, Jaco Du Toit†, and Tobi Awodumila*
*African Institute for Mathematical Sciences, South Africa, §University of Cape Town, South Africa, †Vodacom, South Africa
Email: {oluwaseyi, tobi}@aims.ac.za, jonathan.shock@uct.ac.za, jacowp357@gmail.com

***Abstract*—Dynamic resource allocation in Open RAN (O-RAN) HetNets presents a complex optimisation challenge under varying user loads. We propose a Near-Real-Time RAN Intelligent Controller (Near-RT RIC) xApp utilising Deep Reinforcement Learning (DRL) to jointly optimise transmit power, bandwidth slicing, and user scheduling. Leveraging real-world network topologies, we benchmark Proximal Policy Optimisation (PPO) and Twin Delayed Deep Deterministic Policy Gradient (TD3) against standard heuristics. Our results demonstrate that the PPO-based xApp achieves a superior trade-off, reducing network energy consumption by up to 70% in dense scenarios while improving user fairness by over 30% compared to throughput-greedy baselines. These findings validate the feasibility of centralised, energy-aware AI orchestration in future 6G architectures.**

***Index Terms*—Resource Allocation, Deep Reinforcement Learning, Heterogeneous Networks.**

I. Introduction

The evolution towards fifth-generation (5G) and the forthcoming sixth-generation (6G) wireless systems is driven by a demand for ubiquitous connectivity and high data rates. This has led to the proliferation of Heterogeneous Networks (HetNets), which overlay traditional macrocells with dense tiers of small cells (e.g., micro, pico, and femto cells) to enhance spectral efficiency and network capacity [1], [2]. However, this architectural complexity introduces challenges in resource allocation (RA). The dense deployment of base stations (BS) creates severe co-tier and cross-tier interference, making the efficient management of spectrum, transmit power, and user association critical for network performance. Optimising these resources is essential not only to maximise throughput but also to ensure fairness and quality of service (QoS) for all users in the network [3], [4].

Traditional RA strategies, relying on classical optimisation or heuristics [5], are inadequate for modern HetNets [6]. These methods depend on simplified, static network models and struggle with the nonconvex, combinatorial nature of joint RA problems. Furthermore, distributed approaches often lack the global view necessary for optimal interference coordination. The emergence of Open Radio Access Networks (O-RAN) addresses this by introducing the Near-Real-Time RAN Intelligent Controller (Near-RT RIC), which enables centralised, data-driven control via xApps [7].

Reinforcement learning (RL) has emerged as a powerful paradigm for this challenge. By learning policies through direct interaction with the environment [8], RL agents adapt to real-time conditions without an explicit model. Recent deep reinforcement learning (DRL) approaches effectively handle the high-dimensional state and action spaces of modern networks [9]–[17], demonstrating superior performance over rule-based methods in tasks ranging from power control to network slicing.

While RL for RA is well-investigated, existing work often relies on simplified synthetic topologies or isolates power control from scheduling. This paper bridges the gap between theoretical DRL and realistic deployment constraints. Our specific contributions include formulating a Near-RT RIC-compatible Markov decision process (MDP) in which a central agent manages power and scheduling using global channel knowledge, justified via O-RAN E2 feedback loops. Second, we implement a simulation environment using real-world BS coordinates to capture realistic interference geometries, unlike purely Poisson Point Process (PPP) models. Finally, we provide a mathematically rigorous derivation of throughput and fairness metrics from continuous RL actions, comparing TD3 and PPO against standard heuristics. Simulation results show that DRL agents outperform heuristic baselines by over $\sim 70\%$ in energy reduction and $\geq 100\%$ in throughput while maintaining better fairness among users. The remainder of this paper is organised as follows: Section II details the system model and problem formulation. Section III describes the DRL algorithms. Section IV presents the experimental setup. Section V discusses the results, and Section VI concludes the paper.

II. System Model

We consider a downlink HetNet operating within an O-RAN architecture. The network consists of a set of BSs, $B = \{1, \ldots, N_B\}$, comprising N_M macro BSs and N_S micro BSs.

This work was supported by the African Institute for Mathematical Sciences (AIMS), South Africa, and the Mastercard Foundation Scholarship.
The work of Tobi Awodumila has received funding from the Google DeepMind Scholarship under AIMS.

These serve a set of user equipments (UEs) $U = \{1, \ldots, N_U\}$ distributed stochastically within the coverage area.

The system is controlled by a centralised Near-RT RIC that hosts an xApp responsible for optimising radio resources at discrete time intervals t (cf Fig. 1).

A. Channel Model and Signal Quality

Let $p_{b,t}$ denote the transmit power of BS b at time t, and $x_{b,u} \in \{0, 1\}$ be the binary association variable, where $x_{b,u} = 1$ if user u is served by BS b.

The wireless channel between BS b and user u accounts for path loss, log-normal shadowing, and fast fading. The received power $P^{rx}_{u,b}$ is given by:

$$P^{rx}_{u,b} = p_{b,t} \cdot H_{b,u} \cdot \zeta_{b,u}(t), \tag{1}$$

where $H_{b,u} = d^{-\eta}_{b,u} 10^{\frac{\xi_{b,u}}{10}}$ represents the large-scale channel gain (distance-dependent path loss with exponent η and shadowing $\xi_{b,u} \sim \mathcal{N}(0, \sigma^2_{sh})$). The term $\zeta_{b,u}(t)$ represents the small-scale Rayleigh fading component, assumed to be unit-mean exponential random variables.

The Signal-to-Interference-plus-Noise Ratio (SINR) for user u associated with BS b is formulated as:

$$\mathrm{SINR_{u,b}(t)} = \frac{\mathrm{p_{b,t} H_{b,u} \zeta_{b,u}(t)}}{\sum_{\mathrm{j} \in \mathcal{B} \setminus \mathrm{b}} \mathrm{p_{j,t} H_{j,u} \zeta_{j,u}(t) + N_0 W}}, \tag{2}$$

where N_0 is the noise spectral density and W is the system bandwidth.

B. Throughput and Energy Metrics

The available bandwidth at BS b, denoted as $W_b \in [0, W]$, is dynamically adjusted to mitigate interference. The scheduler at BS b allocates a fraction $\phi_{u,b}(t)$ of W_b to user u, such that $\sum_{u \in U_b} \phi_{u,b}(t) \leq 1$. The achievable data rate for user u is given by the Shannon capacity:

$$R_u(t) = \sum_{b \in \mathcal{B}} x_{b,u} \cdot \phi_{u,b}(t) W_b(t) \log_2 \left(1 + \mathrm{SINR_{u,b}(t)}\right). \tag{3}$$

We strictly define the network energy consumption $E_{\mathrm{net}}(t)$ as the sum of radiated power:

$$E_{\mathrm{net}}(t) = \sum_{b \in B} p_{b,t}. \tag{4}$$

To quantify user fairness, we utilise Jain's Fairness Index $\mathcal{J}(t)$, defined over the rate vector $R(t) = [R_1(t), \ldots, R_{N_U}(t)]$:

$$\mathcal{J}(\mathbf{R}(t)) = \frac{\left(\sum_{u=1}^{N_U} R_{u(t)}\right)^2}{N_U \sum_{u=1}^{N_U} R_u(t)^2}. \tag{5}$$

C. Optimisation Problem

The objective is to find a joint policy π for power control $\mathbf{p}$, bandwidth slicing $\mathbf{W}$, and scheduling weights $\boldsymbol{\phi}$ that maximises a multi-objective utility function over a horizon T. This creates a non-convex, combinatorial optimisation problem:

$$\max_{\mathbf{p}, \mathbf{W}, \boldsymbol{\phi}} \quad \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \left[\omega_1 \sum_u R_u(t) + \omega_2 \mathcal{J}(t) - \omega_3 E_{\mathrm{net}}(t) \right] \tag{P1}$$

$$\text{s.t.} \quad 0 \leq p_{b,t} \leq P_{\max}, \quad \forall b \in B$$
$$0 \leq W_b(t) \leq W, \quad \forall b \in B$$

Direct solution of (P1) is intractable due to the coupling of interference in SINR (2) and the continuous-discrete nature of variables.

D. MDP Formulation for O-RAN xApp

To solve (P1), we formulate the problem as a MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. The agent (xApp) interacts with the environment (E2 nodes) as follows:

State Space $\mathcal{S}$: The state s_t aggregates global network observables available at the RIC:

$$s_t = \left\{ \mathbf{p}_{t-1}, \{\mathbf{I}^{\mathrm{est}}_u\}_{u \in U}, \mathbf{L}_{\mathrm{geo}} \right\}, \tag{6}$$

where $\mathbf{p}_{t-1}$ is the previous power allocation, $\mathbf{I}^{\mathrm{est}}_u$ is the estimated interference measurement from UE channel quality indicator (CQI) reports, and $\mathbf{L}_{\mathrm{geo}}$ encapsulates the fixed topology geometry.

Hierarchical Action Space $\mathcal{A}$: To bridge the timescale gap between RIC (approx. 10ms - 1s) and medium access control (MAC) scheduling (1ms), the agent learns high-level policy parameters rather than instantaneous scheduling decisions. The action vector $a_t \in [-1, 1]^{3N_B}$ consists of normalised values mapped to physical quantities:

- Power Control ($\hat{p}_b$): Scaled to $p_{b,t} \in [P_{\min}, P_{\max}]$.
- Bandwidth Slice ($\hat{w}_b$): Scaled to $W_b(t) \in [0, W]$.
- User Priority Weight ($\hat{psi}_{b,u}$): This scalar influences the local scheduler. The actual resource fraction $\phi_{u,b}$ is derived via a softmax function to ensure validity and differentiability:

$$\phi_{u,b}(t) = \frac{e^{(\tau \cdot \hat{\psi}_{b,u})}}{\sum_{k \in U_b} e^{(\tau \cdot \hat{\psi}_{b,k})}}, \tag{7}$$

where τ is a temperature parameter. This effectively enables the RL agent to bias the local proportional fair scheduler towards specific users (e.g., cell-edge) to enforce fairness.

Reward Function $\mathcal{R}$: The reward r_t is a direct scalarisation of the objective in (P1):

$$r_t = \alpha \frac{\sum R_u(t)}{R_{\max}} + \beta \mathcal{J}(t) - \kappa \frac{\sum p_{b,t}}{N_B P_{\max}}, \tag{8}$$

where coefficients α, β, κ prioritise throughput, fairness, and energy efficiency, respectively. Normalisation terms ensure numerical stability during gradient descent.

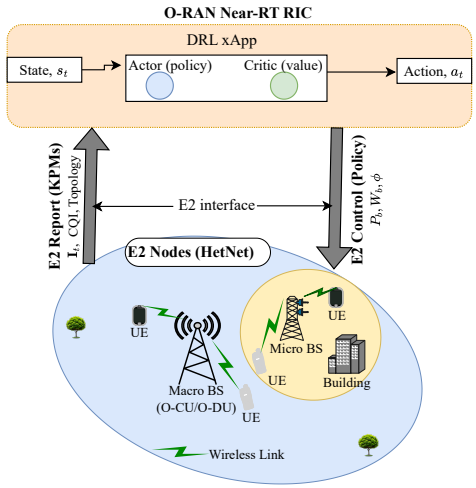

Fig. 1: The proposed O-RAN compliant architecture. The Deep RL agent operates as an xApp within the Near-RT RIC, collecting Key Performance Measurements (KPMs) via the E2 interface to construct the state s_t and issuing optimizing control policies a_t to the Macro and Micro E2 nodes.

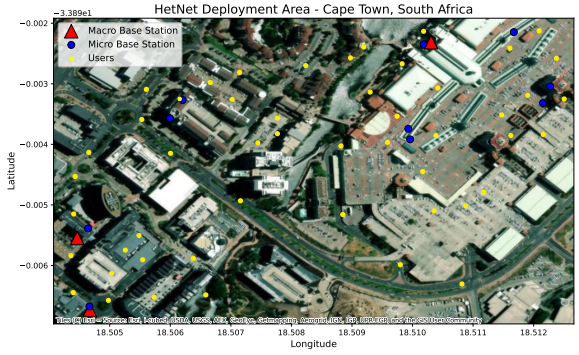

Fig. 2: Satellite image of the deployment area. Source: Esri GIS software for mapping and spatial analysis.

III. DRL Algorithms

The RA problem formulated in Section II is characterised by a high-dimensional state space and a continuous action space (for transmit power and bandwidth). This renders DRL algorithms, such as Deep Q-Networks (DQN), which are restricted to discrete actions, unsuitable. Consequently, we turn to actor-critic and policy-gradient methods, which are designed for continuous control. While Deep Deterministic Policy Gradient (DDPG) is a natural starting point, it is known to suffer from instability and overestimation of Q-values. We therefore select two state-of-the-art algorithms that address these challenges: TD3, which directly mitigates DDPG's shortcomings, and PPO, renowned for its robustness and stable training performance.

A. Twin Delayed Deep Deterministic Policy Gradient (TD3)

TD3 is an off-policy, model-free algorithm that builds upon DDPG by introducing several key modifications to enhance stability and performance. It learns a deterministic policy (the actor) that maps states to actions, and a Q-function (the critic) that estimates the action-value function. The three core innovations of TD3 are:

Clipped Double Q-Learning: To combat the overestimation bias of the critic, TD3 employs two independent critic networks, Q_{θ_1} and Q_{θ_2}. When computing the target value for the Bellman update, it takes the minimum of the two critics' predictions, yielding a more conservative and stable target:

$$y = r + \gamma \min_{i=1,2} Q_{\theta_i'}\left(s', \pi_{\mu'}(s') + \epsilon\right) \tag{9}$$

Where μ' and θ' are the parameters of the target networks, and the noise ϵ is for target policy smoothing.

Delayed Policy Updates: The actor network (π_μ) is updated less frequently than the critic networks. This allows the critic's Q-value estimates to converge and stabilise before being used to update the actor, leading to more reliable policy improvements.

Target Policy Smoothing: Noise is added to the target action during the target Q-value calculation. This helps regularise the policy, making it less likely to exploit narrow peaks in the value function, resulting in a smoother policy landscape.

B. Proximal Policy Optimisation (PPO)

PPO is an on-policy actor-critic algorithm known for its balance between sample efficiency and ease of implementation. Unlike TD3, PPO learns a stochastic policy, $\pi_\theta(a|s)$. Its key feature is a novel surrogate objective function that constrains the size of policy updates, preventing destructive, large changes during training. The core of PPO is the clipped surrogate objective function, which modifies the standard policy gradient objective. It uses the ratio between the new policy and the old policy, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, to measure the policy change. The objective is:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] \tag{10}$$

Where $\hat{A}_t$ is an estimate of the advantage function (often computed using Generalised Advantage Estimation, GAE), and ϵ is a small hyperparameter that defines the clipping range. This objective clips the probability ratio, which discourages policy updates that move $r_t(\theta)$ far from 1, thereby ensuring more stable training. See Appendix for TD3 and PPO training loop.

IV. Experimental Scenarios and Setup

A. Simulation Environment and Topology

We developed a custom O-RAN-compliant simulation environment to evaluate the proposed RIC xApp.

Topology: The network layout is instantiated using real-world BS geospatial data from a telecom operator in Cape Town, South Africa. The dataset comprises $N_M = 3$ macro BSs and $N_S = 10$ micro BSs. While BS locations are fixed to preserve realistic interference geometries, $N_U = 50$ users are randomly distributed within the deployment polygon at the start of each episode to ensure the policy generalises across spatial distributions. Fig. 2 shows the satellite view used to derive the layout. Colors in all figures follow the evaluation convention: Macro BS (red), Micro BS (blue), Users (yellow).
Channel Parameters: The channel propagation follows the model defined in Section II-A. We set the path-loss exponent $\eta = 3.5$ and the log-normal shadowing standard deviation $\sigma_{sh} = 8$ dB to model a dense urban environment. The small-scale fading $\zeta_{b,u}(t)$ is modelled as independent and identically distributed (i.i.d.) Rayleigh fading (unit mean exponential gain), updated every Transmission Time Interval (TTI) to capture fast channel variations. The system bandwidth is $W = 20$ MHz, and thermal noise density is $N_0 = -174$ dBm/Hz.

B. Action Mapping and Hyperparameters

The RL agent's normalised actions $a_t \in [-1, 1]$ are mapped to physical resources as follows:
Power: Transmit power $p_{b,t}$ is scaled linearly. We set $P_{\max}$ to 46 dBm for macro BSs and 30 dBm for micro BSs, with a dynamic range of 20 dB.
Scheduling: The softmax temperature parameter is set to $\tau = 1.0$, allowing the agent to smoothly transition between strict priority scheduling and round-robin behaviour.

C. Training and Evaluation

We train TD3 and PPO agents over 1000 episodes with a horizon of $T = 1000$ steps per episode. The reward function weights in (8) are tuned via grid searches to $\alpha = 1.0$, $\beta = 2.0$, and, $\kappa = 0.5$, prioritising equitable service coverage. We compare the DRL agents against three baselines: **(1) Greedy OFDMA (G-OFDMA)**: assigns RB to the user with the best SINR, **(2) Interference Pricing (IP-PC)**: reduces power based on neighbour feedback, and **(3) Proportional Fair (PF-EQ)**: standard baseline for fairness.

D. Performance Metrics

To evaluate the proposed O-RAN xApp against the baselines, we assess the trained policies on a hold-out test set using the following physical key performance indicators (KPIs):

Average Per-User Throughput ($\bar{R}_{avg}$): This metric quantifies the mean data rate available to an individual user, serving as a primary indicator of Quality of Service (QoS). It is calculated by averaging the instantaneous rates (3) across all users and time steps:

$$\bar{R}_{avg} = \frac{1}{T_{eval} \cdot N_U} \sum_{t=1}^{T_{eval}} \sum_{u=1}^{N_U} R_u(t) \quad \text{[Mbps]}. \tag{11}$$

Average Fairness Index ($\bar{\mathcal{J}}$): To ensure the policy does not maximise throughput by starving cell-edge users, we report the time-averaged Jain's fairness index. This corresponds to the stability of the fairness objective defined in (5):

$$\bar{\mathcal{J}} = \frac{1}{T_{eval}} \sum_{t=1}^{T_{eval}} \mathcal{J}(\mathbf{R}(t)). \tag{12}$$

A value closer to 1 indicates an equitable distribution of resources among all users, regardless of their channel conditions.

Network Energy Consumption ($\bar{E}_{\text{net}}$): We evaluate the environmental impact of the xApp by measuring the average aggregate transmission power of the network, derived from (4):

$$\bar{E}_{\text{net}} = \frac{1}{T_{eval}} \sum_{t=1}^{T_{eval}} \sum_{b=1}^{N_B} p_{b,t} \quad \text{[Watts]}. \tag{13}$$

Lower values indicate that the agent successfully learns to mitigate interference by reducing power rather than simply increasing it.

Average Reward ($\bar{r}$): For the DRL agents, we track the cumulative reward per episode to analyse convergence behaviour and sample efficiency. This serves as a holistic metric of how well the agent balances the conflicting objectives of throughput, fairness, and energy, as defined in (8).

V. Performance Comparison and Discussion

We evaluate the proposed O-RAN xApps (PPO and TD3) against heuristics across four topological scenarios: Dense Urban ($10N_S, 3N_M$, high interference), Sparse Suburban ($3N_M$ only), Hotspot (users cluster near N_S), and Mixed (random N_S plus uniform users). The analysis focuses on the trade-offs between the conflicting objectives defined in Section II

A. Throughput - Energy Trade-off

The trade-off between spectral efficiency and green networking is evident in the Dense Urban and Hotspot scenarios (Figs. 3a and 3c). G-OFDMA achieves a competitive average throughput, but at the cost of maximum energy consumption (normalised $\bar{E}_{\text{net}} \approx 0.95 - 1.0$). Ignoring inter-cell interference forces all BSs to transmit at peak power. IP-PC successfully minimises energy ($\bar{E}_{\text{net}} \approx 0.15$) but results in the lowest user throughput due to overly aggressive power back-off in response to interference pricing.

PPO xApp strikes an optimal balance. In the Dense Urban scenario, PPO achieves a $\sim 70\%$ reduction in energy consumption compared to G-OFDMA while maintaining superior per-user throughput ($\bar{R}_{avg}$). This confirms the agent effectively learns to utilise the "silent" periods and power control actions ($\hat{p}_b$) to mitigate cross-tier interference, maximising the SINR rather than just the signal power.

B. Fairness and QoS Assurance

This is a critical requirement for 6G O-RAN to ensure equitable Service Level Agreements (SLAs). Across all scenarios, G-OFDMA yields poor fairness ($\bar{\mathcal{J}} < 0.25$), indicating that cell-edge users are starved to maximise the sum-rate of cell-centre users. PPO demonstrates superior fairness management,

achieving a Jain's Index of $\approx 0.65 - 0.75$ in all topologies. Notably, in the Mixed Scenario (Fig.3d), PPO improves fairness by over 300% compared to G-OFDMA and 100% compared to IP-PC. While PF-EQ is designed for fairness, it lacks the interference coordination capabilities of the global xApp, resulting in significantly lower aggregate throughput than the DRL agents.

C. Learning Convergence and Computational Complexity

Fig. 3e illustrates the training trajectory of the DRL agents. TD3 exhibits high sample efficiency, converging rapidly within the first 3.5×10^5 steps. However, it suffers from instability and performance degradation in later stages, likely due to overestimation of values in the complex interference landscape. In contrast, PPO demonstrates a stable, monotonic ascent, ultimately achieving a significantly higher mean reward.

Fig. 3f quantifies the computational overhead. The heuristic baselines (G-OFDMA, IP-PC) operate in near-real-time ($< 10^{-1}$s per batch). The DRL inference times are orders of magnitude higher, with PPO being the most computationally intensive. However, the inference latency remains within the $10ms - 1s$ window, validating the deployment of these agents as Near-RT RIC xApps rather than real-time MAC schedulers.

The results indicate that while TD3 offers faster initial deployment, PPO is the superior candidate for the RIC xApp. It provides a robust policy that maximises aggregate utility, successfully protecting cell-edge users (high fairness) and reducing the carbon footprint (low energy use) without compromising network capacity.

VI. Conclusion

In this paper, we addressed the resource orchestration problem in O-RAN HetNets by comparing PPO and TD3-based xApps. Our findings, based on realistic network topologies, reveal that while TD3 converges faster initially, PPO achieves a significantly higher overall reward by learning more effective policies for energy conservation and user fairness. This highlights a critical trade-off: TD3 is a sample-efficient algorithm suitable for rapid adaptation, whereas PPO's methodical exploration yields a more globally optimal policy for performance-critical, energy-constrained environments. Future work will focus on extending this framework to distributed multi-agent scenarios and incorporating the effects of high-speed user mobility.

Acknowledgement

We thank Claude Formanek for his initial assistance with the algorithm design.

Appendix

We provide pseudo code for both TD3 (Algorithm 1) and PPO (Algorithm 2) for the resource allocation problem.

Algorithm 1 TD3 for Resource Allocation Optimisation

Initialise actor π_μ, critics Q_{θ_1}, Q_{θ_2}, and their target networks $\pi_{\mu'}, Q_{\theta_1'}, Q_{\theta_2'}$ and replay buffer $\mathcal{D}$.
for each training step **do**
  Select action with exploration noise: $a = \pi_\mu(s) + \mathcal{N}(0, \sigma)$.
  Store (s, a, r, s') in $\mathcal{D}$ and sample a minibatch from $\mathcal{D}$
  Compute target action with smoothed noise: $a' \leftarrow \pi_{\mu'}(s') + \text{clip}\left(\mathcal{N}(\prime, \sigma), -c, c\right)$.
  Compute target Q-value: $y = r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', a')$
  Update critics θ_i by minimising Huber/MSE loss: $\mathcal{L}(\theta_i) = \left(Q_{\theta_i}(s, a) - y\right)^2$.
  if step is a policy update step **then**
    Softly update all target networks: $\theta' \leftarrow \tau\theta + (1 - \tau)\theta', \mu' \leftarrow \tau\mu + (1 - \tau)\mu'$.
  end if
end for

Algorithm 2 PPO for Resource Allocation Optimisation

Initialise actor-critic network parameters θ.
for each iteration **do**
  Collect a set of trajectories by running policy $\pi_{\theta_{\text{old}}}$ in the environment for T timesteps.
  Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$ (using GAE).
  for a fixed number of epochs **do**
    Optimise the surrogate objective on the collected data via stochastic gradient ascent: $\theta \leftarrow \theta + \alpha \nabla_\theta L^{\text{CLIP}}(\theta)$
  end for
  $\theta_{\text{old}} \leftarrow \theta$.
end for

References

[1] X. Yongjun, G. Guan, G. Haris, and A. Fumiyuki, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.

[2] A. H. Faeq, H. M. Nour, D. Kaharudin, H. E. Binti, S. Nurhizam, Q. Faizan, A. Khairul, and N. Q. Ngoc, "A survey on resource management for 6G heterogeneous networks: Current research, future trends, and challenges," *Electronics*, vol. 12, no. 3, 2023. [Online]. Available: https://doi.org/10.3390/electronics12030647

[3] A. Bharat, T. M. Amine, M. Marco, and M. Gabriel-Miro, "A comprehensive survey on radio resource management in 5G hetnets: Current solutions, future trends and open issues," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2495–2534, 2022. [Online]. Available: https://doi.org/10.1109/COMST.2022.3207967

[4] D. Ather, R. Kler, Z. T. Baig, G. P. Babu, A. Rastogi, and N. Ahmed, *6G Networks: Pioneering Advanced Communication Techniques for Call Centers and Beyond.* CRC Press, 2025. [Online]. Available: https://doi.org/10.1201/9781003583127-12

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. [Online]. Available: https://web.stanford.edu/~boyd/cvxbook/

[6] A. Mughees, M. Tahir, M. A. Sheikh, A. Amphawan, Y. K. Meng, A. Ahad, and K. Chamran, "Energy-efficient joint resource allocation in 5G hetnet using multi-agent parameterized deep reinforcement learning," *Physical Communication*, vol. 61, p. 102206, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874490723002094

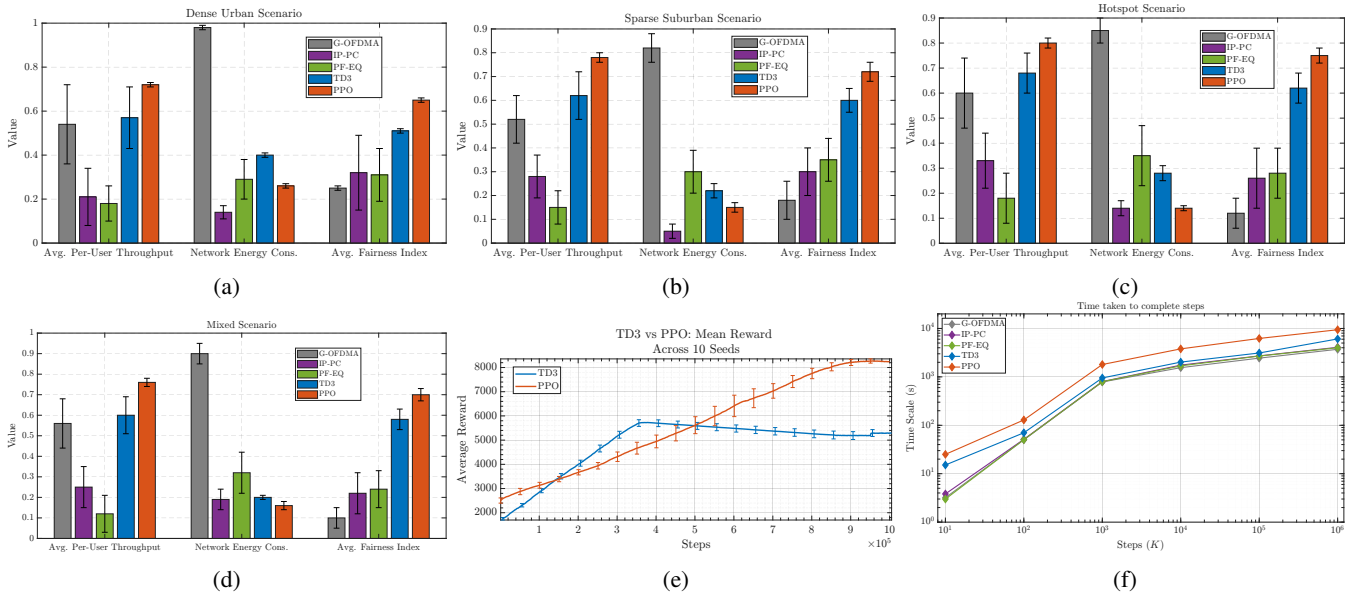

Fig. 3: Comprehensive performance evaluation of the DRL-based O-RAN xApps against heuristic baselines. (a) Dense Urban Scenario: PPO significantly reduces network energy consumption by mitigating cross-tier interference. (b) Sparse Suburban Scenario: PPO achieves near-optimal fairness comparable to PF-EQ while maintaining high throughput. (c) Hotspot Scenario: DRL agents successfully balance load in high-traffic clusters. (d) Mixed Scenario: demonstrating policy robustness to randomised user distributions. (e) Mean reward convergence over 1M steps; PPO demonstrates superior stability compared to TD3. (f) Computational time complexity; the xApp inference latency remains within the Near-RT RIC tolerance window ($10ms - 1s$). Error bars represent the 95% confidence interval.

[7] O. Giwa, M. Adewole, T. Awodumila, and P. Aderinto, "The LLM as a network operator: A vision for generative AI in the 6g radio access network," in *NeurIPS 2025 Workshop: AI and ML for Next-Generation Wireless Communications and Networking*, 2025. [Online]. Available: https://openreview.net/forum?id=81mgAfsFJv

[8] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglu, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *NIPS Deep Learning Workshop 2013*, 2013.

[10] van Hado Hasselt, G. Arthur, and S. David, "Deep reinforcement learning with double q-learning," ser. AAAI'16. AAAI Press, 2016, p. 2094–2100. [Online]. Available: https://doi.org/10.48550/arXiv.1509.06461

[11] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. H. J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," *arXiv preprint, arXiv:1812.05905v2*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1812.05905

[12] A. Archi, H. A. Saadi, and S. Mekaoui, "Applications of deep reinforcement learning in wireless networks-a recent review," in *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM)*, vol. 1, 2023, pp. 1–8.

[13] D. Tian, "An intelligent optimization method for wireless communication network resources based on reinforcement learning," *Journal of Physics: Conference Series*, 2023. [Online]. Available: https://doi.org/10.1088/1742-6596/2560/1/012036

[14] X. Chi, Z. Peifeng, Y. Haibin, and L. Yonghui, "D3qn-based multi-priority computation offloading for time-sensitive and interference-limited industrial wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 9, pp. 13 682–13 693, 2024. [Online]. Available: https://doi.org/10.1109/TVT.2024.3387567

[15] J. Park and W. Na, "Application of mac protocol reinforcement learning in wireless network environment," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, 2024, pp. 730–731.

[16] K. Olayemi, M. Van, S. McLoone, Y. Sun, J. Close, N. M. Nyat, and S. McIlvanna, "A twin delayed deep deterministic policy gradient algorithm for autonomous ground vehicle navigation via digital twin perception awareness," *arXiv preprint, arXiv:2403.15067v1*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.15067

[17] S. Shalini, N.Kopperundevi, R.Rajkumar, A. Radhika, M. Gopianand, and M. Ram, "Decentralized machine learning for dynamic resource optimization in wireless networks using reinforcement learning," *Journal of Electrical Systems*, 2024. [Online]. Available: https://doi.org/10.52783/jes.2539