SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Deep Reinforcement Learning for MC-MIMO-NOMA: A Corrected Benchmark Reveals No Gains over Classical Beamforming and Power Control

Gurudevi A. Hiremath[1], Shankar D. Nawale[2]

[1]SKN College of Engineering, Korti, Pandharpur, India and PAH Solapur University, Solapur, Maharashtra, India.
E-mail: gurudevihiremath17@gmail.com
[2]N. B. Navale Sinhgad College of Engineering, Solapur, Maharashtra, India.

ABSTRACT
Deep reinforcement learning (DRL) is increasingly proposed for joint beamforming, power control and user scheduling in multi-cell multiple-input multiple-output non-orthogonal multiple access (MC-MIMO-NOMA) systems, frequently with reported gains over classical optimization. This paper is a reproducibility study and corrected benchmark. Beginning from a public DRL framework for this problem, we (i) identify and fix simulation and modelling defects that render its results uninterpretable—most critically an unknown-interference power set roughly 40 dB above the received signal, which makes the per-user signal-to-interference-plus-noise-ratio (SINR) constraint infeasible in every tested configuration, an energy-detector receiver in place of a linear combiner, an inverted successive-interference-cancellation (SIC) term, a non-functional reconfigurable-intelligent-surface model, and an inconsistent policy log-likelihood; (ii) show, on the corrected and QoS-feasible environment, that end-to-end DRL fails to learn transmit-power minimization—learned power is essentially flat in the QoS threshold even in the interference-free single-user regime, whereas the analytically optimal power varies by roughly 20 dB, and this persists when training is extended by 6.7×; and (iii) study a hybrid that pairs beamforming directions with a classical optimal power-control inner loop (the Foschini–Miljanic fixed point) and greedy admission. The hybrid recovers the expected power–QoS–fairness tradeoffs and guarantees QoS for all served users, but DRL-learned directions underperform every classical baseline we test— maximum ratio transmission (MRT), regularized zero-forcing (RZF), weighted MMSE (WMMSE), and signal-to-leakage-plus-noise-ratio (SLNR) beamforming—serving as few as one third of the users SLNR serves at the same QoS, across interference and load sweeps and over multiple seeds. Our results do not support the prevailing claim that DRL outperforms classical methods for this problem. We release a corrected, QoS-feasible, statistically rigorous benchmark and delineate the conditions under which learning could plausibly add value.

Keywords: MIMO-NOMA, beamforming, power control, user scheduling, reinforcement learning, benchmarking, SLNR, WMMSE.

1. Introduction

The application of deep reinforcement learning (DRL) to physical-layer resource allocation has expanded rapidly. Numerous works report that learned policies match or exceed classical optimization for power control, beamforming and scheduling in NOMA and MIMO systems [1]–[4]. The appeal is clear: the joint beamforming–scheduling problem in multi-cell MIMO-NOMA is a mixed-integer nonlinear program (MINLP) that is NP-hard in general, and a trained network can, in principle, amortize the cost of near-optimal decisions into a single fast forward pass. The credibility of any "DRL beats classical" claim, however, rests on three foundations that are easy to overlook and hard to verify from a paper alone: (a) a correctly modelled environment in which the quality-of-service (QoS) constraints are actually feasible, so that "performance" is meaningful;
(b) strong, fairly tuned classical baselines rather than weak strawmen; and (c) statistically rigorous, reproducible evaluation with multiple seeds and honest error reporting. The DRL literature at large has documented how fragile empirical conclusions can be when these are neglected [13]. Wireless DRL is not immune.

This study: This paper is a reproducibility study. Taking a representative, publicly described DRL framework for MC-MIMO-NOMA [1] as a starting point, we set out to reproduce its central claim—that a generative-AI-enhanced primal-dual proximal policy optimization agent minimizes transmit power better than conventional baselines. We were unable to do so. Diagnosing the failure led first to a series of simulation and modelling defects and then, after correcting them, to a different and—we argue—more useful set of conclusions about when learning helps for this problem.

Our central finding is negative but constructive: on a corrected, QoS-feasible environment and against strong classical baselines, end-to-end DRL does not learn the power-minimization objective, and DRL-learned beamforming directions are dominated by classical designs. The one configuration that works—a hybrid pairing directions with classical optimal power control and admission—derives its performance from its

classical components. We release everything required to reproduce these results.
Contributions: The authors identify and fix defects that made the original benchmark uninterpretable. The most consequential is an unknown-interference power set 40 dB above the received signal, which renders the SINR constraint infeasible in every configuration; we also replace an energy-detector receiver with an MMSE interference-rejection combiner (MMSE-IRC), correct an inverted SIC term and an inter-cell interference mask, fix an imperfect-CSI evaluation error, and remove a non-functional RIS model. After correction, the received SINR lies in a realistic range and QoS becomes attainable.
On the corrected environment, DRL does not learn transmit-power minimization: learned power is flat in the QoS threshold γmin even in the interference-free single-user regime, while the analytically optimal power varies by 20 dB. The gap is not a training-budget artifact—it persists when training is extended to 1000 episodes. The authors attribute the failure to high per-step reward variance from fast-fading channels combined with advantage normalization.
This study follows a hybrid combining beamforming directions with the classical Foschini–Miljanic optimal power-control fixed point and greedy admission. It recovers the expected power–QoS–fairness tradeoffs and guarantees QoS for all served users. Against four classical direction designs (MRT, RZF, WMMSE, SLNR), DRL-learned directions serve the fewest users at every threshold, across QoS, interference and load sweeps and multiple seeds.
The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the system model and the MMSE-IRC SINR. Section 4 documents the reproduction attempt and corrections. Sections 5 and 6 present the DRL negative result and the hybrid with its baselines and derivations. Section 7 details the reproducible setup. Sections 8–9 present results and integrity findings. Sections 10 and 11 discuss and conclude.

2. Related Work

2.1. DRL for NOMA and MIMO Resource Allocation

DRL has been applied across wireless resource management, including power control in interference channels [3], continuous beamforming design [5], and joint subchannel/power allocation in NOMA [4]. Proximal policy optimization (PPO) [6], soft actor-critic, and deterministic policy gradient methods are common choices. The framework we build upon [1] augments a primal-dual PPO agent with a convolutional inverted-transformer feature extractor and generative pre-training for joint beamforming and scheduling in MC-MIMO-NOMA, reporting gains over a PPO baseline. Our study revisits that class of claims under a corrected environment and stronger baselines.

2.2. Classical Beamforming and Power Control

Linear precoding for the multi-user MIMO downlink is well studied. Regularized zero-forcing (RZF) [7] trades off interference nulling against noise enhancement; signal-to-leakage-plus-noise-ratio (SLNR) beamforming [8] admits a closed-form generalized-eigenvector solution that suppresses leakage to all other users; and weighted MMSE (WMMSE) [9] iteratively maximizes weighted sum rate and is a standard strong baseline. Given fixed beamforming directions and a linear receiver, minimizing transmit power subject to SINR targets reduces to a linear power-control problem with a well-known optimal fixed point and feasibility condition [10], [11]; feasibility and admission for SINR-constrained systems are classical [12]. We use these as the power-control and direction components of the hybrid and as baselines.

2.3. Reproducibility in (Wireless) DRL

The broader machine-learning community has documented severe reproducibility problems in DRL: sensitivity to seeds, implementation details and evaluation protocol [13]. Wireless DRL studies inherit these risks and add domain-specific ones—most importantly, whether the simulated environment is physically consistent and QoS-feasible. To our knowledge, corrected, QoS-feasible, multi-seed benchmarks with strong classical baselines are rare for MC-MIMO-NOMA. This paper contributes one.

2.4. Positioning of This Work

Table I compares representative recent learning-based studies for NOMA and multi-user MIMO beamforming/power/scheduling against this work. Two patterns are evident. First, baselines are usually weak or learning-only: several papers compare only against other DRL methods [17], [19], and those with a classical anchor typically use a single one (ZF [15], MVDR [23], SCA [22], fractional programming [16]); a full classical suite spanning MRT, RZF, WMMSE, coordinated multi-cell WMMSE and SLNR is essentially absent, and strong QoS-feasible power minimizers [20] are rarely used as the yardstick. Second, multi-seed statistics and released, corrected environments are rare: nearly all report single-run curves, and none of the RL-vs-classical comparisons combine released code, a corrected/verified environment, and multi-seed error bars. Reported gains also frequently hide that the “win” is over other DRL/heuristics, or that DRL only reaches 85–90% of a classical optimum [15], [16], or that QoS is only softly satisfied [18], [22]. To our knowledge this work is the first to combine all of the distinguishing attributes in the last row of Table I.

3. System Model

The authors consider the downlink of an M -cell MIMO-NOMA network. Base station (BS) m has Nt antennas and serves up to K candidate users, each with Nr antennas. Let Hm,k ∈ CNr×Nt be the channel from BS m to its user k, and Gi,m,k ∈ CNr×Nt the cross channel from interfering BS i ≠ m to user (m, k). Channels follow a geometric multipath model with L paths, where βm,k is the path gain (3GPP-style path loss), is the gl∼CN(0,1) and and at, ar are uniform-linear-array responses. Parameters are listed in Table II.

3.1. Receiver, SIC and SINR

User (m, k) applies a linear combiner um,k CNr . With NOMA successive interference cancellation, users in a cell are ordered by channel gain, π(1) ≥ π(2) ≥· · · , and user k cancels the messages of weaker co-cell users it can decode, suffering residual interference only from stronger, un-cancelled co-cell users, i.e. those with π(j) < π(k). The realized SINR is given in (2) at the top of the next page, where σ2 is thermal noise and σ2 the unknown-source interference power. MMSE-IRC combiner. Let sm,k = Hm,kdm,k be the effective desired signature and Rm,k the interference-plus-noise covariance at user (m, k):

$$\mathbf{H}_{m,k} = \sqrt{\frac{\beta_{m,k}}{L}} \sum_{\ell=1}^{L} g_\ell \, \mathbf{a}_r(\theta_\ell^r)\mathbf{a}_t(\theta_\ell^t)^H, \qquad (1)$$

$$\gamma_{m,k} = \frac{p_{m,k}\,|\mathbf{u}_{m,k}^H \mathbf{H}_{m,k}\mathbf{d}_{m,k}|^2}{\sum_{j:\pi(j)<\pi(k)} p_{m,j}|\mathbf{u}_{m,k}^H \mathbf{H}_{m,k}\mathbf{d}_{m,j}|^2 + \sum_{i\neq m}\sum_{j} \alpha_{i,j}\, p_{i,j}|\mathbf{u}_{m,k}^H \mathbf{G}_{i,m,k}\mathbf{d}_{i,j}|^2 + \sigma_I^2\|\mathbf{u}_{m,k}\|^2 + \sigma_n^2\|\mathbf{u}_{m,k}\|^2}$$

$$\mathbf{R}_{m,k} = \sum_{j:\pi(j)<\pi(k)} p_{m,j}\,\mathbf{h}_{m,k,j}\mathbf{h}_{m,k,j}^H + \sum_{i\neq m}\sum_{j} \alpha_{i,j} p_{i,j}\mathbf{g}_{i,m,k,j}\mathbf{g}_{i,m,k,j}^H + (\sigma_I^2 + \sigma_n^2)\mathbf{I}, \qquad (3)$$

$$\gamma_{m,k} = p_{m,k}\,\mathbf{s}_{m,k}^H \mathbf{R}_{m,k}^{-1}\mathbf{s}_{m,k}. \qquad (4)$$

TABLE I. Comparison of recent learning-based studies for NOMA / MU-MIMO resource allocation vs. this work. MC/SC: multi-/single-cell. Baseline strength: strong = WMMSE/SLNR/ global-optimal; one = a single classical method; DRL-only = compared only to learning/heuristics. Sd: multi-seed statistics; Cd: code released; CSI: imperfect/correlated CSI studied; QoS: per-user QoS guaranteed (Y/soft/N)

Study	System	Method	Obj.	Classical baseline	Sd	Cd	CSI	QoS
Mismar et al. [14]	MC MISO	DQN (E2E)	SINR/rate	link-adapt. (weak)	N	Y	N	N
Feng & Clerckx [15]	SC massive MIMO	multi-agent DRL	rate	ZF (one)	N	N	Y (age)	N
Wang et al. [16]	SC mmWave MIMO-NOMA	DQN+DDPG hy-brid	sum-rate	frac. prog. (one)	N	N	N	soft
Sun et al. [17]	MC NOMA (SISO)	multi-agent DQL	energy eff.	DRL-only	N	N	N	soft
Li & Liu [18]	SC multicast MISO	transformer L2O	power (QoS)	SDR/SCA (strong)	N	N	N	soft
Kim & Ankireddy [19]	MC (MISO/MIMO)	offline RL (BCQ)	rate	DRL-only	N	Y	N	N
Norouzi et al. [20]	SC MIMO-NOMA	classical (BB)	power (QoS)	global-opt (self)	–	N	N	Y
Shi et al. [21]	SC MISO NOMA+RIS	LSTM-DDPG (re-cur.)	Success ratio	DRL-only	N	N	∼	soft
Zhang et al. [22]	MIMO RSMA (sat.)	constrained SAC	energy eff.	SCA (one)	N	N	Y (err.)	soft
Al Janaby et al. [23]	MC MU-MIMO	multi-agent DQN	SINR	MVDR (one)	N	N	N	N
This work	MC MIMO-NOMA	hybrid + DRL critique	power (QoS)	MRT, RZF, WMMSE, CoWMMSE, SLNR	Y	Y	Y (d.+c.)	Y

with hm,k,j =Hm,kdm,j and gi,m,k,j =Gi,m,kdi,j. The MMSE-IRC combiner um,k = R−1m,kSm,k maximizes the output SINR, which then admits the compact form. To see (4), note the SINR after combining is γm,k = pm,k|uH sm,k|2/(uH Rm,ku); by the generalized Rayleigh quotient this is maximized by u ∝ R−1 sm,k, giving the maximum value pm,k sH R−1 sm,k. Equation (4) is what our environment computes; it replaces the physically inconsistent energy detector of the original code (Section IV), under which the desired and each interfering term were passed through different matched filters.

TABLE II: Simulation Parameters

Parameter	Value	Parameter	Value
Cells M	7	Users/cell K	8 (4–12)
BS antennas Nt	32	User antennas Nr	4
Multipaths L	6	Cell radius	800 m
Path loss	30.6+36.7 log10d	Noise σ2	−99 dBm
Pmax	46 dBm	Interf. (INR)	σI (swept)
γmin	0–15 dB	Seeds	up to 15

3.2. Problem Formulation

The objective is minimum total transmit power subject to per-user QoS and per-BS budget. Problem (5) is an MINLP: the binary α couple to continuous (d, p) through the SINR, and admission (choosing which users to serve when not all can meet QoS) is intrinsic.

4. Reproduction Attempt And Corrections

The authors unable to reproduce the headline "DRL minimizes power" result of the baseline framework. Diagnosis revealed defects that, individually and jointly, invalidate the original results. Table III summarizes them; we expand on the most consequential or exceed classical.

(D1) Infeasible SINR by construction. With near-optimal MRT beamforming at full power, the received signal was 74 dBm while thermal noise was 99 dBm, a healthy 25 dB SNR; however, the unknown-interference power was 35 dBm—about 40 dB above the received signal. Consequently, the per-user SINR was 40 to 60 dB in every configuration and the QoS constraint could never be met, regardless of power or beamforming. Reported "transmit power" therefore reflected a reward balance, not constrained optimization. We reference the unknown interference to the noise floor as an interference-to-noise ratio (INR), σ2 =σ2 10INR/10 restoring feasible SINR (5–28 dB, Table IV).

TABLE III: Summary of Defects and Corrections

ID	Defect	Correction
D1	Unknown interference ∼40 dB above signal ⇒ SINR infeasible everywhere	INR referencing to noise floor; feasible SINR restored
D2	Energy-detector "receiver"	MMSE-IRC combiner (4)
D3	Inverted SIC residual interference term	Correct index; gain-based SIC order
D4	Inter-cell mask on victim cell; CSI mismatch absent	Gate by interfering BS; design on estimate, realize on truth
D5	RIS-to-user channel identically zero	RIS excluded pending faithful model

TABLE IV: Worst served-user SINR ceiling at Pmax (MRT beamforming), corrected environment: MMSE-IRC vs the energy detector

Scheduled users/cell	Energy detector (D2)	MMSE-IRC (this work)
1	16.2 dB	28.1 dB
2	1.3 dB	12.6 dB
3	−5.6 dB	4.3 dB

(D2) Energy-detector receiver. The original SINR used Hw 2 for the desired term and, independently, for each interferer—equivalent to applying a different matched filter to every signal, which no single physical receiver can realize and which inflates the desired-to-interference ratio. We replace it with the MMSE-IRC combiner (4). This raises the achievable worst-user SINR ceiling substantially (Table IV), making NOMA pairing feasible.
(D3) Inverted SIC. Residual intra-cell interference was summed over weaker (decoded-later) users rather than stronger un-cancelled users; we correct the index in (2) and order users by pre-beamforming channel gain to remove a circular dependence between the decoding order and the beamformer being optimized.
(D4) Inter-cell mask and imperfect CSI. Inter-cell interference was gated by the victim cell's schedule instead of the interfering BS's; and imperfect-CSI experiments used the same errored channel for both beamforming and SINR, so no mismatch penalty existed. We gate by $\alpha_{i,j}$ and, under imperfect CSI, beamform on the estimate while realizing SINR on the true channel.
(D5) Non-functional RIS. The reconfigurable-intelligent-surface-to-user channel was identically zero, so the RIS contributed nothing and the reported RIS "gains" were run-to-run noise. We exclude RIS from the corrected benchmark pending a faithful model.

5. End To End Drl Does Not Learn Power Minimization

The authors evaluate a primal-dual PPO agent [6] on the corrected environment. To give learning every advantage, we additionally correct three issues in the released agent that would otherwise handicap it: (i) the policy log-likelihood applied the tanh squashing correction at sampling but not in the update, biasing the importance ratio—we make them consistent; (ii) the value head was sigmoid-bounded to [0, 1] while returns were 103, rendering the baseline useless—we make it unbounded; and (iii) we provide an explicit per-user power action, so the policy controls power directly rather than through the magnitude of MKNt 2 beamforming coefficients. Admission is fixed to a feasible top-k set to isolate power learning. Despite these advantages, learned power is essentially flat in γmin. In the interference-free single-user-per-cell regime—the easiest, purely power-limited case—trained power is 33.9 dBm at γmin = 0 and 33.9 dBm at γmin = 12 dB, whereas the analytically optimal power rises from 6.7 to 20 dBm (Fig. 1). The agent over-provisions by 15–27 dB and does not track the threshold. The behaviour is identical in multi-user regimes.

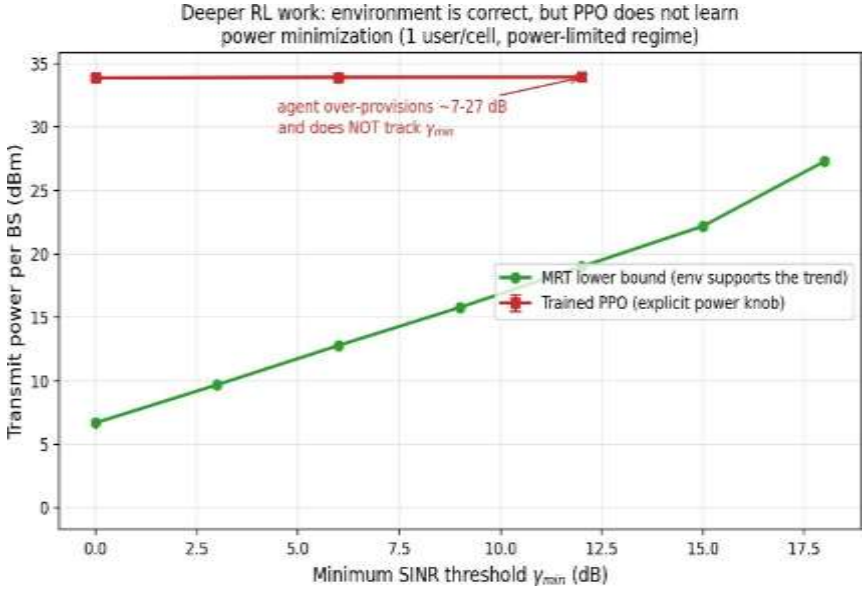

Figure 1: Single-user (power-limited) regime: the corrected environment supports the power–QoS trend (analytic lower bound, green), but trained PPO over-provisions and does not track γmin (red, mean ± std over seeds).

The environment regenerates independent fast-fading channels every step, so per-step reward variance from channel realizations dominates the smooth effect of the power action combined with PPO advantage normalization, which removes the absolute power-cost scale, the power-minimization gradient is buried. Consistent with this, the result is invariant to the fixes above and to the action parameterization, indicating a property of the formulation rather than a tuning artifact.

6. A Qos-Guaranteeing Hybrid And A Fair Comparison

The failure above motivates decoupling what optimization solves well (power) from what learning might help with (directions and admission).

6.1. Optimal Power Control for Fixed Directions

Fix the unit-norm directions {dm,k} and a served set $\boldsymbol{S} = \{(m,k): \alpha_{m,k}=1\}$and adopt the fixed MRC combiner $\mathbf{u}_{m,k}$

= $\mathbf{s}_{m,k}/|\mathbf{s}_{m,k}|$ With the combiner fixed, the gains $G_i = |\mathbf{u}^H \mathbf{s}_i|^2$ and couplings $F_{i,j} = |\mathbf{u}^H \mathbf{c}_{i,j}|^2$ no longer depends on power, so SINR of user $i \in S$ is the linear fractional form. The QoS contstraint γi ≥ γmin rearranges to $p_i \geq (\gamma_{min}/G_i)\sum_{j\neq i} p_j F_{i,j} + \eta_i)$;stacking over S yields.

$$\mathbf{p} \geq \mathbf{DFp} + \mathbf{D}\eta, \quad \mathbf{D} = \mathrm{diag}\left(\frac{\gamma_{min}}{G_i}\right), \qquad (6)$$

where Gi = uH si 2 is the desired gain, Fi,j = uH ci,j 2 the interference coupling (ci,j the effective interference channel from user j to i, respecting SIC), and ηi = σ2 + σ2 . The component-wise minimum-power solution meets all constraints with equality [10]. which is feasible (positive, finite) if and only if the spectral radius ρ(DF) < 1. When (7) is infeasible or violates a per-BS budget, admission must remove users.

$$\mathbf{p}^\star = (\mathbf{I} - \mathbf{DF})^{-1}\mathbf{D}\eta, \qquad (7)$$

6.2. Greedy Admission

We adopt the classical greedy rule: solve (7); if infeasible or over budget, drop the costliest served user (largest p⋆, or lowest gain when ρ ≥ 1) and re-solve, never dropping a cell's last user. Algorithm 1 summarizes the hybrid. By construction, every served user meets γmin.

Algorithm 1 Hybrid: directions + optimal power + admission

Input: directions $\{\mathbf{d}_{m,k}\}$, threshold $\gamma_{\min}$
$\mathcal{S} \leftarrow \{(m,k)\}$ (all users)
while $\mathcal{S} \neq \emptyset$ **do**
  build $\mathbf{D}, \mathbf{F}, \eta$ over $\mathcal{S}$ (MRC combiner)
  if $\rho(\mathbf{DF}) < 1$ **then**
    $\mathbf{p}^\star \leftarrow (\mathbf{I} - \mathbf{DF})^{-1}\mathbf{D}\eta$
    if $\mathbf{p}^\star \geq 0$ and per-BS budgets hold **then**
      return $\{\mathbf{w}_{m,k} = \sqrt{p^\star_{m,k}}\,\mathbf{d}_{m,k}\}$, $\mathcal{S}$
    end if
  end if
  drop the costliest user from $\mathcal{S}$
end while

6.3. Beamforming-Direction Methods

We compare five ways to choose the directions fed to Algorithm 1.

- MRT: dm,k is the dominant eigenvector of HHmk,Hm,k as (interference-agnostic).
- RZF [7]: per-BS regularized zero-forcing, dm,k ∝ (AmAH + λI)−1am,k, where Am = [am,1, . . . , am,K] stacks the dominant per-user effective channels of cell m and λ = (σ2 +σ2)K/Pmax; suppresses intra-cell leakage.
- SLNR [8]: the dominant generalized eigenvector.
- WMMSE [9]: per-cell iterative weighted-MMSE

$$\mathbf{d}_{m,k} = \arg\max_{\|\mathbf{d}\|=1} \frac{\mathbf{d}^H \mathbf{H}^H_{m,k}\mathbf{H}_{m,k}\mathbf{d}}{\mathbf{d}^H(\sigma^2\mathbf{I} + \mathbf{L}_{m,k})\mathbf{d}},$$
$$\mathbf{L}_{m,k} = \sum_{k'\neq k} \mathbf{H}^H_{m,k'}\mathbf{H}_{m,k'} + \sum_{i\neq m}\sum_{j} \mathbf{G}^H_{m,i,j}\mathbf{G}_{m,i,j}, \qquad (8)$$

beamformers, directions extracted after convergence.

- DRL: directions from the trained PPO policy (power set by Algorithm 1, so the reward reflects served-users-at-minimum-power).

6.4. Complexity

Each admission iteration solves an S S system and an eigenvalue check, O(S3) with S = MK; with at most S drops the per-decision cost is O(S4) worst case, dominated in practice by the first feasible solve. SLNR/RZF cost O(MK N 3) (closed-form); WMMSE adds a modest number of O(N 3) iterations. DRL inference is a single forward pass but, as shown, yields inferior directions.

TABLE V: PPO Agent Hyperparameters

Parameter	Value	Parameter	Value
Actor/critic net	MLP 2562	Optimizer	Adam
Learning rate	10-3	Discount γ	0.99
PPO clip ϵ	0.2	PPOepochs	4
Steps/episode	40	Batch	Full rollout
Value head	Unbounded	Log-prob	Tanh

Power action	Per-user	Entropy bonus	10-2
Episodes	150 (1000) - for the plateau run	Seeds	3

TABLE VI. Hybrid (MRT directions + Algorithm 1), corrected env, mean over 15 seeds.

γmin (dB)	Power/BS (dBm)	# served/56	Jain	QoS met
0	26.5	39.6	0.89	100%
3	28.9	35.2	0.94	100%
6	31.6	29.9	0.95	100%
9	32.2	25.0	0.97	100%
12	32.4	20.2	0.99	100%

7. Reproduction Experimental Setup

All experiments use the parameters of Table II on the corrected environment. Classical baselines are evaluated over 12–15 independent channel seeds; DRL over 3 training seeds (150 episodes) with an additional 1000-episode run to test the training-budget hypothesis. Power is reported per BS in dBm; "served" counts users meeting γmin under the MMSE-IRC SINR (4); fairness is Jain's index over served-user rates. Every metric is computed by a single source-of-truth analysis script, and all figures/tables are regenerated from released JSON outputs. We report mean and standard deviation across seeds and never claim variance bands that were not computed. The pre-correction numbers reported for comparison—the energy-detector ceilings in Table IV and the infeasible-regime signal/interference levels in Section IV—are reproducible from the released code via legacy toggles (a one-line script). The PPO agent's hyperparameters are given in Table V; all are released in configuration files.

8. RESULTS

8.1. The Hybrid Recovers the Expected Physics

With MRT directions and Algorithm 1 (Table VI, Fig. 2, 15 seeds), transmit power rises monotonically with γmin (26.5 → 32.4 dBm), the served set tightens (39.6 → 20.2), Jain's fairness is high and rising (0.89 → 0.99), and 100% of served users meet QoS at every threshold—none of which end-to-end DRL achieved. This is the power–QoS–fairness tradeoff the problem should exhibit.

8.2. DRL Directions Underperform all Classical Baseline

Using the same power-control and admission inner loop, we compare DRL directions to MRT, RZF, WMMSE and SLNR (Table VII, Fig. 3). Every method guarantees 100% QoS; the differentiator is how many users can be served. DRL serves the fewest at every threshold—about 40% of MRT and roughly one third of the strongest baseline SLNR (16.1 vs 48.4 at γmin = 0). The gap is not a training-budget artifact: extending PPO to 1000 episodes (6.7×) leaves served users at ~16, plateaued by episode 50 (Fig. 4).

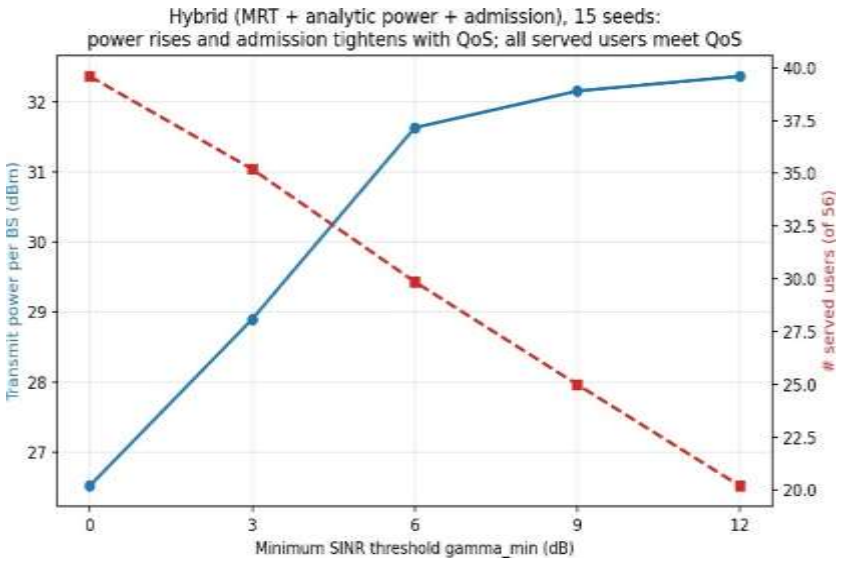

TABLE VII. Served users (of 56) meeting QoS: learned vs classical Directions, All with the same Optimal Power Control + admission (all achieve 100% QoS; higher is better). Classical: mean over 15

seeds; DRL (PPO): mean±std over 3 seeds

γmin (dB)	DRL (PPO)	MRT	RZF	WMMSE	SLNR
0	16.1±0.1	39.6	41.7	41.5	48.4
6	10.7±0.1	29.9	31.5	30.3	36.6
12	7.2±0.0	20.2	23.3	22.0	26.1

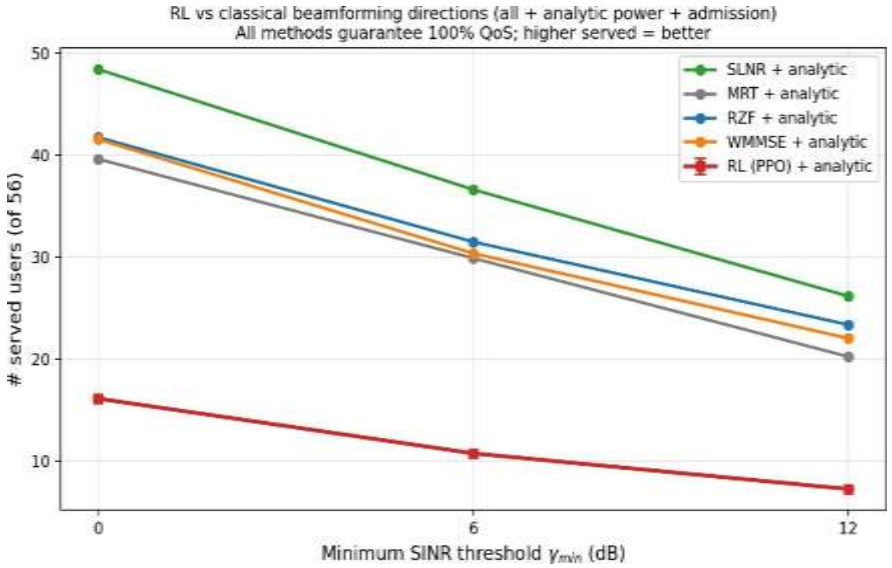

Figure 3: Served users vs QoS threshold for learned (PPO) and classical directions, all sharing the optimal power-control + admission inner loop. DRL is dominated by all classical baselines; SLNR is strongest

8.3. Robustness Across Interference and Load

The conclusions hold across operating conditions (Fig. 5, 12 seeds, γmin = 6 dB). As the unknown-interference INR rises from −10 to +10 dB, all methods maintain QoS with a gradual reduction in served users, SLNR remaining strongest (~36 vs MRT ~32). As the user pool K grows from 4 to 12, admission exploits the larger pool and served users rise (SLNR 22.1 → 42.4; MRT 19.7 → 39.2), with SLNR dominant throughout. In every condition the hybrid guarantees 100% QoS, SLNR is the best direction method, and DRL would lie below the MRT curve.

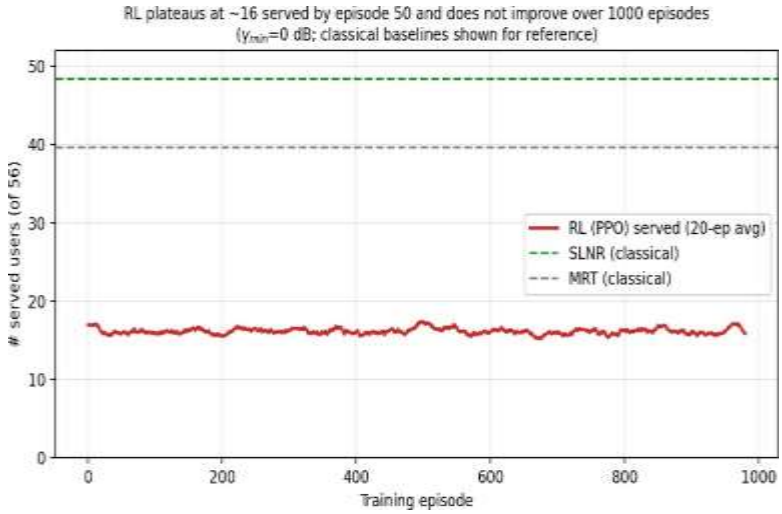

Figure 4: Training trajectory at γmin=0. PPO plateaus at ~16 served users by episode 50 and does not improve through 1000 episodes, remaining far below the classical baselines (dashed).

9. DATA INTEGRITY PITFALL IN THE ORIGINAL BENCHMARK

As guidance for the community, we record integrity problems found in the original released artifacts, each of which a trustworthy benchmark must avoid: (i) the headline "7.35% improvement" of the proposed agent over PPO is not reproducible from the released data and is sign-reversed (the PPO baseline is at least as good); (ii) figure captions claim " 1 standard deviation over three seeds" but the code runs a single seed and contains no error-bar computation; (iii) reported Jain fairness indices have no computational provenance, as per-user rates

are not logged; (iv) all transmit-power figures apply a double dBm conversion (10 log10( ) + 30 to values already in dBm), inflating and compressing the power axis. Our released benchmark re-implements every metric, logs per-user rates, computes genuine multi-seed error bars, and centralizes analysis in a single script.

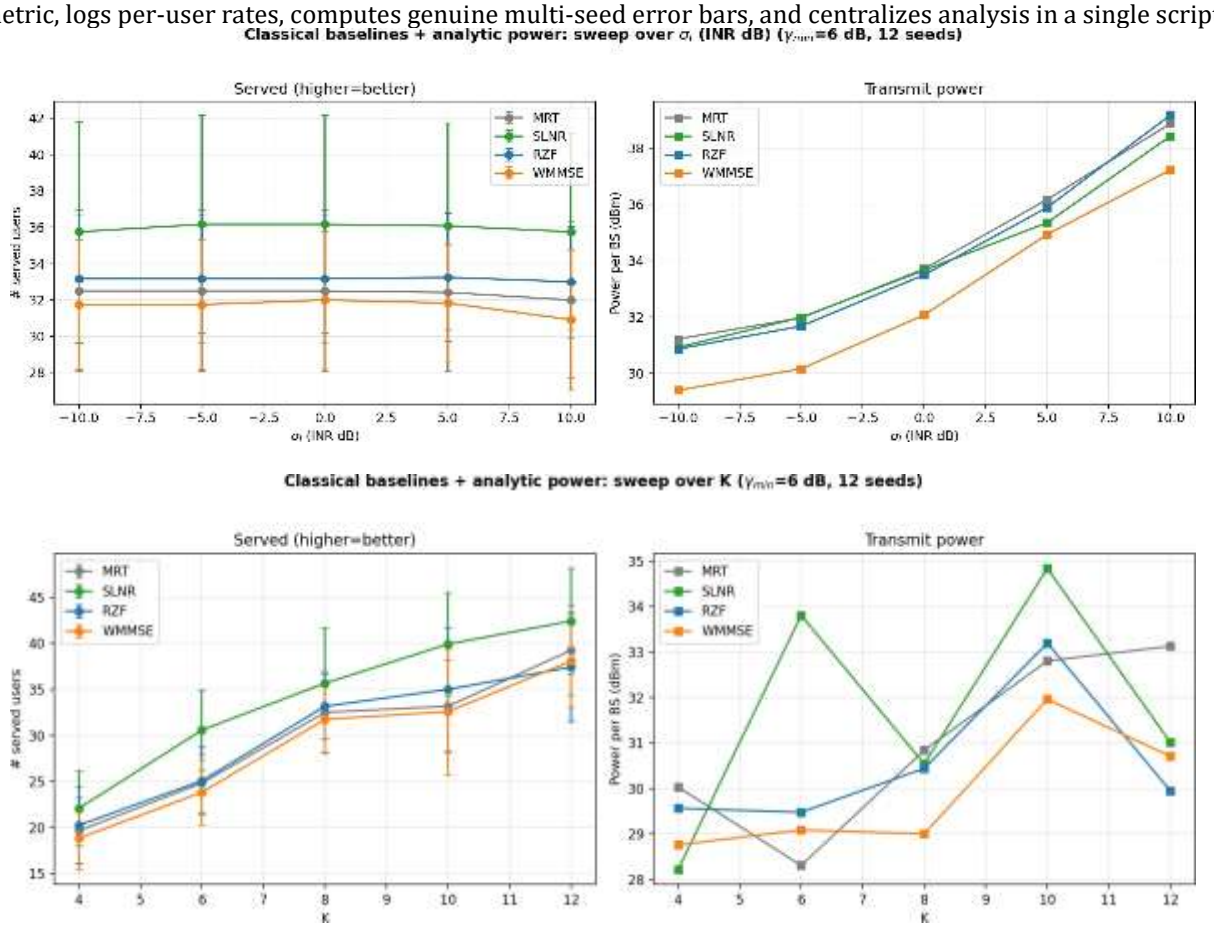

Figure 5: Served users (mean ± std, 12 seeds) for classical directions + optimal power vs unknown-interference INR (left) and number of users K (right), at γmin = 6 dB. SLNR dominates across all conditions; all methods guarantee

10. Discussion

Our results do not support "DRL beats classical" for MC-MIMO-NOMA beamforming and power control as commonly formulated. Two mechanisms explain this. First, power control is solved optimally and cheaply by classical means (7); learning the power magnitude adds nothing and is hard to train under fast-fading reward variance. Second, for directions, learning must beat closed-form designs that already exploit the channel structure; our evidence shows DRL does not beat even MRT, let alone SLNR/RZF/WMMSE. The authors see three plausible niches, each requiring the corrected benchmark and strong baselines to validate: (i) admission/scheduling under combinatorial intercell coupling, where classical methods are heuristic and the action is naturally discrete; (ii) settings with temporally correlated channels (so a recurrent/attention policy can exploit memory), unlike the i.i.d. per-step channels here—though Appendix C shows that even here a GRU policy does not beat a closed-form MMSE predictor, because the latter already captures Gauss–Markov temporal structure optimally; and (iii) amortized inference under tight latency budgets, where a trained network trades a small optimality gap for constant-time decisions—but this must be measured against the latency of (7), not assumed. Of these, (iii) appears the most promising, as it concerns computation rather than decision quality. The DRL representative is primal-dual PPO; other architectures might differ, though the failure mechanism (reward variance vs. smooth power effect) is architecture-agnostic and the direction gap to SLNR is large. Our WMMSE and RZF are per-cell; a fully coordinated multi-cell WMMSE could serve more users still, which would only strengthen the conclusion. Results are for i.i.d. block-fading; correlated channels are future work.

11. Conclusion

A reproducibility study of DRL for multi-cell MIMO-NOMA found that the original environment was infeasible by construction and that, once corrected, end-to-end DRL does not learn power minimization and DRL-learned beamforming directions underperform all classical baselines tested (MRT, RZF, WMMSE, SLNR), across QoS, interference and load sweeps and multiple seeds, and regardless of training budget. A classical hybrid—beamforming directions with optimal power control and greedy admission—recovers the expected physics

and guarantees QoS. We release the corrected, QoS-feasible, statistically honest benchmark. The broader lesson is that claims of learned-over-classical superiority in physical-layer resource allocation demand feasible environments, strong baselines, and reproducible, multi-seed evaluation; we hope the released benchmark helps enforce that standard.

Appendix A

Coordinated Multi-Cell WMMSE

The per-cell WMMSE baseline of Section VI treats intercell interference as background noise. We additionally implement a coordinated multi-cell WMMSE that accounts for the leakage each beamformer causes to users in all cells. Table VIII reports served users over an independent 15-seed draw (hence small differences from Table VII, e.g. SLNR at γmin = 0 reads 45.1 here vs 48.4 there—within seed variation, as the per-seed served count has standard deviation 6). Coordinated WMMSE (CoWMMSE) is competitive with the other classical methods and, like all of them, far exceeds DRL (Table VII); under i.i.d. fading SLNR remains strongest. All methods again guarantee 100% QoS.

TABLE VIII. Served users (of 56) including coordinated multi-cell WMMSE (15 seeds, i.i.d. fading). All methods achieve 100% QoS.

γmin (dB)	MRT	RZF	WMMSE	CoWMMSE	SLNR
0	40.2	39.9	38.9	35.3	45.1
6	28.9	30.6	28.3	31.6	35.1
12	21.3	24.8	23.5	23.9	26.9

Appendix B

Correlated-Fading Robustness

The main results use i.i.d. block fading. We assess robustness to transmit spatial correlation via an exponential Kro-necker model, H HR1/2 with [Rt]i,j = r|i−j|, r [0, 1). Fig. 6 reports served users vs r at γmin = 6 dB (12 seeds). Two observations. First, all classical methods maintain 100% QoS at every correlation level—the hybrid's QoS guarantee is unaffected. Second, the ranking shifts with correlation: served users degrade as r grows (fewer spatial degrees of freedom), but coordinated WMMSE degrades least and becomes the strongest method at high correlation (r = 0.9: 31.8 served vs SLNR's 25.7 and MRT's 21.7), because inter-cell coordination is more valuable when per-link spatial selectivity is reduced. DRL remains non-competitive throughout. Exploiting temporal correlation via recurrent or attention-based policies—the one regime where learning has an information advantage over memoryless snapshot optimizers—is left as future work.

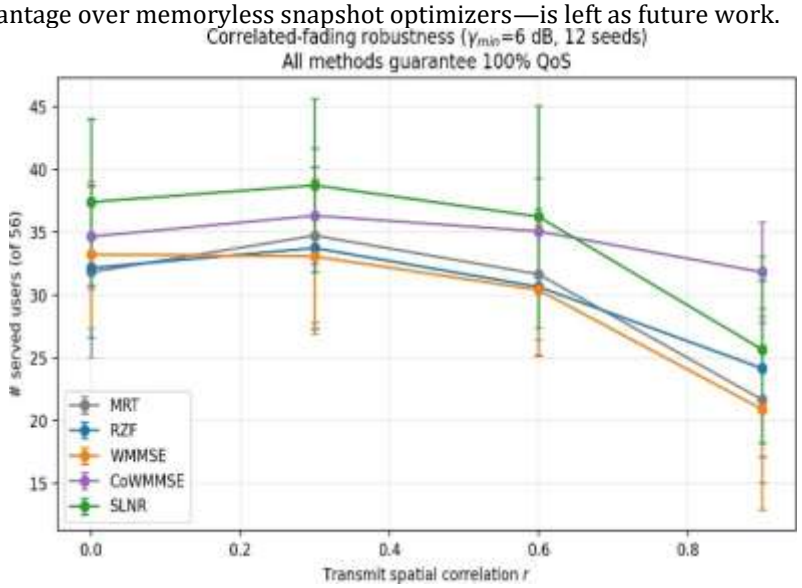

Figure 6: Served users vs transmit spatial correlation r (γmin = 6 dB, 12 seeds). All methods guarantee 100% QoS; coordinated WMMSE is most robust and leads at high correlation

Appendix C

Temporal Correlation and Recurrent RL

The main experiments use i.i.d. block fading and perfect current CSI, under which memory cannot help a policy (the current channel is fully observed). The authors therefore test the one regime in which learning has a genuine information advantage over memoryless snapshot optimizers: temporally correlated fading with imperfect (delayed) CSI, where a policy with memory could learn to track/predict the channel. Even in this best case for learning, the recurrent policy (12.6 3.0 served@QoS) does not beat classical methods with MMSE

prediction (19.3–19.4), and—tellingly—the GRU memory adds nothing over the memoryless policy (12.6 vs 12.4) as shown in figure 7. The reason is fundamental: for a Gauss–Markov channel the MMSE predictor ρ H(t 1) is closed-form and statistically optimal, so a learned recurrent predictor has no residual temporal information to exploit. The classical baseline already captures the temporal structure optimally. This reinforces the paper's conclusion: even granted its most favorable regime and a fair recurrent architecture, DRL does not outperform classical methods here.

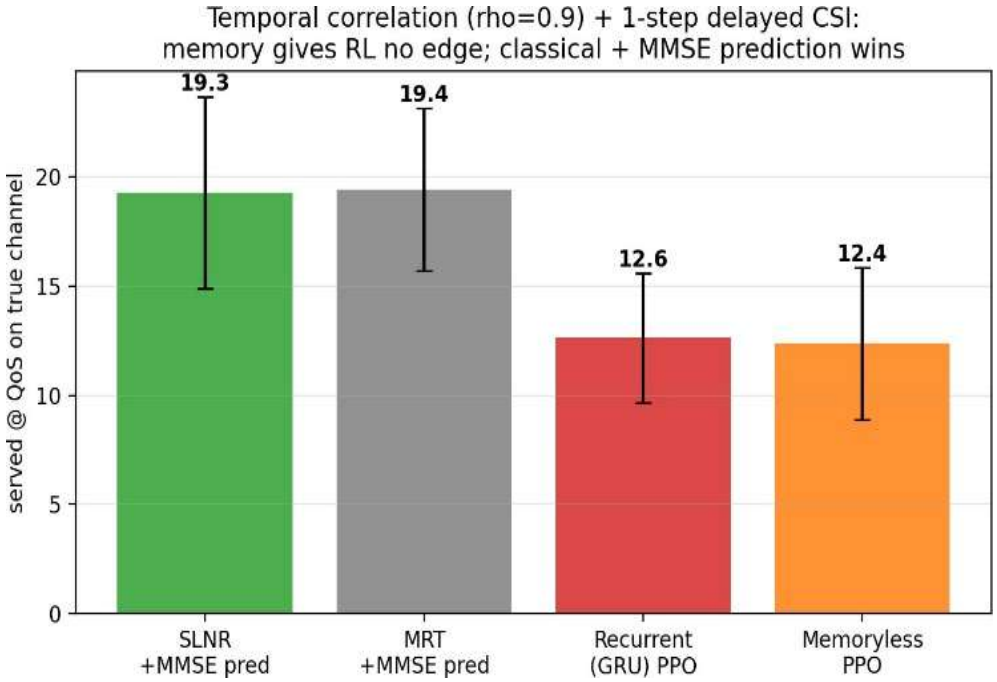

Figure 7: Temporally correlated fading (ρ = 0.9) with one-step delayed CSI: served users meeting QoS on the true channel. Recurrent (GRU) PPO does not beat classical methods with the closed-form MMSE predictor, and memory yields no gain over the memoryless policy

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

The authors declare that no funding received to conduct this research.

Acknowledgments

The authors acknowledge the technical support provided by PAH Solapur University, Solapur in conducting this research work.

References

1. A. B. M. Adam et al., "Generative AI-driven reinforcement learning for beamforming and scheduling in multi-cell MIMO-NOMA systems," Phys. Commun., vol. 72, 2025.
2. N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," IEEE Commun. Surveys Tuts., vol. 21, no. 4, pp. 3133–3174, 2019.
3. Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," IEEE J. Sel. Areas Commun., vol. 37, no. 10, pp. 2239–2250, 2019.
4. C. He et al., "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," IEEE J. Sel. Areas Commun., vol. 37, no. 10, 2019.
5. Z. Zhang et al., "Deep reinforcement learning for beamforming in multiuser MISO systems," IEEE Trans. Wireless Commun., 2020.
6. J. Schulman et al., "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.
7. C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part I: Channel inversion and regularization," IEEE Trans. Commun., vol. 53, no. 1, pp. 195–202, 2005.
8. M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," IEEE Trans. Wireless Commun., vol. 6, no. 5, pp. 1711–1721, 2007.
9. Q. Shi et al., "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," IEEE Trans. Signal Process., vol. 59, no. 9, pp. 4331–4340, 2011.
10. G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its

convergence," IEEE Trans. Veh. Technol., vol. 42, no. 4, pp. 641–646, 1993.
11. J. Zander, "Performance of optimum transmitter power control in cellular radio systems," IEEE Trans. Veh. Technol., vol. 41, no. 1, pp. 57–62, 1992.
12. A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," IEEE Trans. Signal Process., vol. 54, no. 1, pp. 161–176, 2006.
13. P. Henderson et al., "Deep reinforcement learning that matters," in Proc. AAAI, 2018.
14. F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," IEEE Trans. Commun., vol. 68, no. 3, pp. 1581–1592, 2020.
15. Z. Feng and B. Clerckx, "Deep reinforcement learning for multi-user massive MIMO with channel aging," arXiv preprint arXiv:2302.06853, 2023.
16. M. Wang et al., "Spectrum-efficient user grouping and resource allocation based on deep reinforcement learning for mmWave massive MIMO-NOMA systems," Sci. Rep., vol. 14, 2024.
17. M. Sun, Y. Zhong, X. He, and J. Zhang, "Channel and power allocation for multi-cell NOMA using multi-agent deep reinforcement learning and unsupervised learning," Sensors, vol. 25, no. 9, art. 2733, 2025.
18. Y. Li and Y.-F. Liu, "HPE Transformer: Learning to optimize multi-group multicast beamforming under nonconvex QoS constraints," IEEE Trans. Commun., vol. 72, no. 9, pp. 5581–5594, 2024.
19. H. Kim and S. K. Ankireddy, "Learning RL-policies for joint beamforming without exploration: A batch constrained off-policy approach," 2023 (arXiv:2310.08660).
20. S. Norouzi, B. Champagne, and Y. Cai, "Joint optimal design of user clustering, beamforming, and power allocation for NOMA," IEEE Trans. Commun., vol. 71, no. 1, pp. 214–228, 2023.
21. Z. Shi et al., "Active RIS-aided EH-NOMA networks: A deep reinforcement learning approach," IEEE Trans. Commun., vol. 71, no. 10, pp. 5846–5861, 2023.
22. Q. Zhang, L. Zhu, Y. Chen, and S. Jiang, "Constrained DRL for energy efficiency optimization in RSMA-based integrated satellite terrestrial network," Sensors, vol. 23, no. 18, art. 7859, 2023.
23. A. Al Janaby, H. Al-Rizzo, and Y. Qassim, "Enhancing spectral efficiency of 6G downlink beamforming via cooperative multi-agent deep reinforcement learning," Sensors, vol. 26, no. 3, art. 950, 2026.