Cooperative Deep Reinforcement Learning for Fair RIS Allocation

1st Martin Mark Zan
Institute of Telecommunications
TU Wien
Vienna, Austria
martin.zan@tuwien.ac.at

2nd Stefan Schwarz
Institute of Telecommunications
TU Wien
Vienna, Austria
stefan.schwarz@tuwien.ac.at

***Abstract*—The deployment of reconfigurable intelligent surfaces (RISs) introduces new challenges for resource allocation in multi-cell wireless networks, particularly when user loads are uneven across base stations. In this work, we consider RISs as shared infrastructure that must be dynamically assigned among competing base stations, and we address this problem using a simultaneous ascending auction mechanism.**

To mitigate performance imbalances between cells, we propose a fairness-aware collaborative multi-agent reinforcement learning approach in which base stations adapt their bidding strategies based on both expected utility gains and relative service quality. A centrally computed performance-dependent fairness indicator is incorporated into the agents' observations, enabling implicit coordination without direct inter-base-station communication.

Simulation results show that the proposed framework effectively redistributes RIS resources toward weaker-performing cells, substantially improving the rates of the worst-served users while preserving overall throughput. The results demonstrate that fairness-oriented RIS allocation can be achieved through cooperative learning, providing a flexible tool for balancing efficiency and equity in future wireless networks.

***Index Terms*—Reconfigurable Intelligent Surfaces, Resource Allocation, Auctions, Reinforcement Learning, Fairness, Multi-Agent Systems.**

I. Introduction

In the evolution toward 6G wireless networks, intelligent resource management has become a central challenge in interference-limited environments. While advances in spectral efficiency and massive antenna systems have significantly improved peak data rates, ensuring fair and reliable service across users and cells remains a key objective, particularly at the cell edge where propagation conditions are poor and competition for shared resources is most pronounced.

While cell-edge performance limitations are addressed by techniques such as coordinated multipoint transmission (CoMP) [1], [2] and cell-free massive MIMO [3], [4], the practical deployment of these approaches is constrained by limited coordination cluster sizes [5], [6]. As a result, full network-wide coordination remains infeasible, and performance degradation persists at the edges of coordination clusters rather than at conventional cell boundaries. The framework considered in this work is therefore complementary to CoMP and cell-free MIMO.

Reconfigurable intelligent surfaces (RISs) have recently emerged as a promising technology to address these challenges by enabling programmable control of the wireless propagation environment [7]. By adjusting the phase responses of nearly passive reflecting elements, RISs provide a cost- and energy-efficient means to enhance desired signal paths and mitigate interference, thereby complementing conventional base station and user equipment capabilities [8], [9].

Despite their potential, practical RIS deployment raises system-level questions regarding placement and efficient allocation among multiple transmitters and users. These challenges are particularly pronounced in multi-cell scenarios, where RISs deployed near cell boundaries can benefit multiple base stations, leading to competition for shared infrastructure.

To address this competition, RISs are modeled as shared resources managed by an independent infrastructure provider and dynamically leased to base stations (BSs) via a market-inspired allocation mechanism. In particular, auction-based allocation provides a scalable and low-complexity alternative to combinatorial optimization approaches [10], while explicitly capturing the strategic interactions among competing base stations. Similar auction formats have been successfully applied in spectrum allocation, where simultaneous ascending auctions enable efficient distribution of scarce resources [11], [12]. An auction-based RIS allocation mechanism was recently studied in [13] for a multi-operator scenario, demonstrating the viability of this approach in RIS-assisted networks.

Building on this framework, reinforcement learning (RL) enables optimized bidding strategies in dynamic and partially observable environments. By learning from repeated auction interactions, RL agents adapt their behavior to target high-value RISs while avoiding inefficient bidding, and have been shown to outperform heuristic approaches in performance–cost trade-offs [14]. Deep RL has been successfully used to coordinate a central resource provider and multiple competing operators in multi-RIS networks in [15]. Auction-based energy markets have been formulated as stochastic games, where reinforcement learning is used to learn effective bidding strategies under uncertainty and competition in [16].

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT4490824.

In contrast to existing work, this paper studies fairness-aware RIS allocation in asymmetric multi-cell scenarios with uneven user distributions. We introduce a performance-dependent fairness indicator into the RL agents' observations, enabling implicit coordination that favors weaker-performing cells when beneficial. A tunable parameter controls the trade-off between total throughput and equitable resource distribution.

Simulation results show that the proposed cooperative multi-agent RL framework yields a more balanced RIS allocation, significantly improving minimum user rates in overloaded cells and reducing the Atkinson inequality index, while maintaining competitive sum-rate performance. We further demonstrate how fairness settings shape agent behavior and system-level outcomes.

Notation: The multi-variate complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{C})$, while the uniform distribution over the interval $[a, b]$ is written as $\mathcal{U}(a, b)$. For a vector $\boldsymbol{x}$, the transpose and Hermitian transpose are $\boldsymbol{x}^T$ and $\boldsymbol{x}^H$, respectively, and the i-th element is denoted by $\boldsymbol{x}[i]$. The Euclidean norm of $\boldsymbol{x}$ is $\|\boldsymbol{x}\|$. The cardinality of a set $\mathcal{X}$ is $|\mathcal{X}|$, and the empty set is denoted by $\emptyset$. The expectation operator is written as $\mathbb{E}[\cdot]$, and the phase of a complex number z is given by $\arg(z)$.

II. System Model

We consider a multi-cell downlink scenario with N_{BS} base stations, serving a total of N_{UE} single-antenna users with the assistance of N_{RIS} reconfigurable intelligent surfaces. Each BS is equipped with M_{BS} antennas, while each RIS consists of M_{RIS} reflecting elements. We consider single-user MIMO, i.e., users are served on orthogonal resources.

A. Channel Model

We consider both direct and RIS-assisted channels between each BS and each user. We assume that the direct BS-user (UE) link is dominated by non-line-of-sight (NLOS) propagation and is strongly shadowed, which motivates the use of RIS-assisted transmission. The direct channel between user u and BS b is denoted by $\boldsymbol{h}_{u,b}^{\mathrm{direct}} \in \mathbb{C}^{M_{\mathrm{BS}} \times 1}$ and modeled as

$$\boldsymbol{h}_{u,b}^{\mathrm{direct}} = \gamma_{u,b} \boldsymbol{g}_{u,b}, \tag{1}$$

where $\gamma_{u,b}$ is the path gain, and $\boldsymbol{g}_{u,b} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ is the fading component. The normalization is chosen such that $\mathbb{E}[\|\boldsymbol{g}_{u,b}\|^2] = M_{\mathrm{BS}}$, yielding $\mathbb{E}[\|\boldsymbol{h}_{u,b}^{\mathrm{direct}}\|^2] = \gamma_{u,b}^2 M_{\mathrm{BS}}$. We consider Rayleigh fading for this link to model strong multipath propagation under NLOS conditions.

For the channel between BS b and RIS r, a strong line-of-sight (LOS) component is assumed, as RISs are typically deployed in locations with good visibility to nearby BSs. This directional LOS channel is represented by the matrix $\boldsymbol{H}_{r,b} \in \mathbb{C}^{M_{\mathrm{RIS}} \times M_{\mathrm{BS}}}$ and modeled as

$$\boldsymbol{H}_{r,b} = \gamma_{r,b} \boldsymbol{a}(\psi_{r,b}) \boldsymbol{a}(\theta_{r,b})^T, \tag{2}$$

where $\gamma_{r,b}$ is the corresponding path gain, $\boldsymbol{a}(\psi_{r,b})$ denotes the RIS array response vector, and $\boldsymbol{a}(\theta_{r,b})$ is the BS array response vector [17]. The angles $\psi_{r,b}$ and $\theta_{r,b}$ describe the angle-of-arrival at the RIS and the angle-of-departure at the BS, respectively. We assume that additional multipath components are negligible for this link.

For the channel between RIS r and user u we consider both, a LOS component and additional multipath components modeled as Rayleigh fading. The overall RIS-user channel, denoted by $\boldsymbol{h}_{u,r} \in \mathbb{C}^{M_{\mathrm{RIS}} \times 1}$, therefore follows a Rician fading model

$$\boldsymbol{h}_{u,r} = \gamma_{u,r} \left(\sqrt{\frac{K_{u,r}}{1 + K_{u,r}}} \boldsymbol{a}(\theta_{u,r}) + \sqrt{\frac{1}{1 + K_{u,r}}} \boldsymbol{g}_{u,r} \right), \tag{3}$$

where $\gamma_{u,r}$ is the path gain, $K_{u,r}$ denotes the Rician K-factor, $\boldsymbol{a}(\theta_{u,r})$ is the RIS array response vector, and $\boldsymbol{g}_{u,r} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ models the NLOS component. The angle-of-departure at the RIS is $\theta_{u,r}$. The previously defined path gains $\gamma_{u,b}, \gamma_{r,b}, \gamma_{u,r}$ depend on both distance and line-of-sight conditions.

Each RIS applies a diagonal phase-shift matrix

$$\boldsymbol{\Phi}_r = \mathrm{diag}\big(e^{j\phi_{r,1}}, \ldots, e^{j\phi_{r,M_{\mathrm{RIS}}}}\big), \tag{4}$$

where the phase shifts are configured to coherently align the LOS components of the RIS-assisted channel. We assume that the random scattering components $\boldsymbol{g}_{u,r}$ vary rapidly over time, which prevents their reliable estimation. Consequently, these components cannot be coherently phase-aligned at the RIS and therefore contribute only incoherently to the received signal. If a RIS is not assigned to the serving BS, its phase shifts are assumed to be random, i.e., $\phi_{r,i} \sim \mathcal{U}(0, 2\pi)$.

The aggregate RIS-assisted channel between BS b and user u, $\boldsymbol{h}_{u,b}^{\mathrm{indirect}} \in \mathbb{C}^{M_{\mathrm{BS}} \times 1}$ is given by

$$\boldsymbol{h}_{u,b}^{\mathrm{indirect}} = \sum_{r=1}^{N_{\mathrm{RIS}}} \left(\boldsymbol{h}_{u,r}^T \boldsymbol{\Phi}_r \boldsymbol{H}_{r,b} \right)^T. \tag{5}$$

The total channel $\boldsymbol{h}_{u,b} \in \mathbb{C}^{M_{\mathrm{BS}} \times 1}$ is then

$$\boldsymbol{h}_{u,b} = \boldsymbol{h}_{u,b}^{\mathrm{direct}} + \boldsymbol{h}_{u,b}^{\mathrm{indirect}}. \tag{6}$$

B. Beamforming Model

We assume strong shadowing of the direct link, such that users are effectively served only via RISs. Consequently, base station beamforming is directed toward RISs and is based solely on the directional channel components represented by the array response vectors of the dominant paths. Rapidly varying NLOS components are not exploited, as they cannot be reliably estimated.

We point beams towards the RISs and we assign power across these beams in a user-specific way. Let $\boldsymbol{f}_{u,d} \in \mathbb{C}^{M_{\mathrm{BS}} \times 1}$ denote the beamforming vector used by the serving base station d for user u, including the power allocation, such that $\mathbb{E}[\|\boldsymbol{f}_{u,d}\|^2] = P_d$.

When a set $\mathcal{R}^{(d)}$ of RISs is assigned to base station d, the beamforming vector is constructed as

$$\boldsymbol{f}_{u,d} = \sqrt{\frac{1}{M_{\mathrm{BS}}}} \sum_{r \in \mathcal{R}^{(d)}} \sqrt{P_{u,r,d}}\, \boldsymbol{a}^*(\theta_{r,d}), \tag{7}$$

where $\boldsymbol{a}(\theta_{r,d})$ is the array response vector at the base station corresponding to the direction of RIS r, $P_{u,r,d}$ denotes the power allocated to user u via RIS r, and $\sum_{r\in\mathcal{R}^{(d)}} P_{u,r,d} = P_d$. The power allocation will be defined in Section III.

If no RIS is assigned to base station d, directional information is unavailable. In this case, the base station applies random Gaussian beamforming for the users it serves.

C. Signal Model

We consider a downlink transmission model in which each BS serves multiple users using orthogonal time-frequency resources. As a result, intra-cell interference is not present. However, inter-cell interference remains present due to simultaneous transmissions from neighboring BSs.

The received signal at user u served by BS d is obtained as

$$y_u = \boldsymbol{h}_{u,d}^T \boldsymbol{f}_{u,d} x_u + \sum_{b\neq d} \boldsymbol{h}_{j_b,b}^T \boldsymbol{f}_{j_b,b} x_j + n_u, \tag{8}$$

where j_b denotes the user which is served by BS b at the same time as user u is served by BS d, and $n_u \sim \mathcal{CN}(0,\sigma_n^2)$ is additive white Gaussian noise. The transmit symbol intended for user u is denoted by x_u and satisfies $\mathbb{E}[|x_u|^2] = 1$.

D. SINR Model

The signal-to-interference-plus-noise ratio (SINR) for user u served by BS d is expressed as

$$\mathrm{SINR}_u^{(d)} = \frac{\left|\boldsymbol{h}_{u,d}^T \boldsymbol{f}_{u,d}\right|^2}{\sigma_n^2 + \sum_{b\neq d} \frac{1}{|\mathcal{U}^{(b)}|} \sum_{j\in\mathcal{U}^{(b)}} \left|\boldsymbol{h}_{u,b}^T \boldsymbol{f}_{j,b}\right|^2}. \tag{9}$$

The users assigned to base station b are denoted by $\mathcal{U}^{(b)}$. For tractability, we replace the instantaneous inter-cell interference by its average over users in neighboring cells, yielding a scheduling-agnostic interference model. This approximation reflects the fact that channel coding spans many resource elements with potentially varying interferers, and the resulting SINR is interpreted as an effective long-term metric.

The achievable downlink rate of user u served by BS d is then given by

$$r_u^{(d)} = \log_2\left(1 + \mathrm{SINR}_u^{(d)}\right). \tag{10}$$

III. SINR and Utility Estimation

In order to evaluate RIS allocations and guide the auction-based resource assignment, each base station requires an estimate of the achievable performance under a given RIS configuration. Since instantaneous channel state information is not available prior to RIS allocation and configuration, we rely on macroscopic channel parameters and asymptotic properties of large antenna arrays to estimate the SINR and the resulting utility.

A. Macroscopic SINR Estimation

We approximate the instantaneous received power terms by their respective expected values. For sufficiently large antenna arrays and RISs, this is justified by the law of large numbers.

The estimated SINR of user u served by base station d is expressed as

$$\hat{\mathrm{SINR}}_u^{(d)} = \frac{p_{d,u} + p_{c,u} + p_{i,u}}{\sigma_n^2 + i_{d,u} + i_{i,u}}, \tag{11}$$

where $p_{d,u}$ denotes the direct signal power, $p_{c,u}$ and $p_{i,u}$ represent the coherent and incoherent RIS-assisted signal components, respectively, and $i_{d,u}$ and $i_{i,u}$ denote direct and RIS-assisted interference. The noise is denoted by σ_n^2.

B. Signal Power Estimation

It is justified to split the signal power from (9) into two parts, direct and indirect powers, since we assume a Rayleigh fading link for the direct path. As the Rayleigh fading components are assumed to be unknown for beamforming and RIS configuration, this implies that the indirect RIS-assisted channels are not coherently combined with the direct NLOS channel.

We will first estimate the direct part. The beamforming vector is statistically independent from the direct channel, because it is matched to the RIS channel. We have:

$$\mathbb{E}\left[\left\|(\boldsymbol{h}_{u,d}^{\text{direct}})^T \boldsymbol{f}_{u,d}\right\|^2\right] = \mathbb{E}\left[\left\|\gamma_{u,d}\boldsymbol{g}_{u,d}^T \boldsymbol{f}_{u,d}\right\|^2\right] =$$

$$\gamma_{u,d}^2 \boldsymbol{f}_{u,d}^H \mathbb{E}\left[\boldsymbol{g}_{u,d}^* \boldsymbol{g}_{u,d}^T\right] \boldsymbol{f}_{u,d} = \gamma_{u,d}^2 \left\|\boldsymbol{f}_{u,d}\right\|^2 = \gamma_{u,d}^2 P_d = p_{d,u},$$

where $\mathbb{E}\left[\boldsymbol{g}_{u,d}^* \boldsymbol{g}_{u,d}^T\right] = \boldsymbol{I}$.

The RIS-assisted signal (the indirect part of the signal from (9)) is the following:

$$\mathbb{E}\left[\left\|(\boldsymbol{h}_{u,d}^{\text{indirect}})^T \boldsymbol{f}_{u,d}\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{h}_{u,r}^T \boldsymbol{\Phi}_r \boldsymbol{H}_{r,d} \boldsymbol{f}_{u,d}\right\|^2\right].$$

We additionally assume asymptotic orthogonality:

$$\boldsymbol{a}(\theta_{r,d})^T \boldsymbol{a}(\theta_{s,d})^* \approx 0, \forall s \neq r, \quad \boldsymbol{a}(\theta_{r,d})^T \boldsymbol{a}(\theta_{r,d})^* = M_{\text{BS}}.$$

The RIS-assisted signal contains both coherent and incoherent components. The coherent component arises from the line-of-sight parts of the RIS-UE channels that can be phase-aligned by the RISs, while the incoherent component is caused by the non-line-of-sight Rayleigh fading contributions, which cannot be coherently combined as they are assumed to be unknown. After substituting the channel and beamforming expressions, the coherent signal over RIS $r \in \mathcal{R}^{(d)}$ is:

$$s_{u,r,d} = \underbrace{\gamma_{u,r}\gamma_{r,d} k_{u,r} M_{\text{RIS}} \sqrt{M_{\text{BS}}}}_{c_{u,r,d}} \underbrace{\sqrt{P_{u,r,d}}}_{m_{u,r,d}},$$

where $k_{u,r} = \sqrt{K_{u,r}/(1+K_{u,r})}$ and $m_{u,r,d}$ is obtained from the power allocation. The total received signal over all allocated RISs is given by:

$$p_{c,u} = \left(\sum_{r\in\mathcal{R}^{(d)}} c_{u,r,d} m_{u,r,d}\right)^2 = \boldsymbol{m}_{u,d}^T (\boldsymbol{c}_{u,d}\boldsymbol{c}_{u,d}^T) \boldsymbol{m}_{u,d}, \tag{12}$$

where for $r_i \in \mathcal{R}^{(d)}$: $\boldsymbol{c}_{u,d} = [c_{u,r_1,d}, \ldots, c_{u,r_{\mathcal{R}^{(d)}},d}]$ and $\boldsymbol{m}_{u,d} = [m_{u,r_1,d}, \ldots, m_{u,r_{\mathcal{R}^{(d)}},d}]$.

We maximize (12) with respect to the power allocation with the constraint $||\boldsymbol{m}_{u,d}||^2 = P_d$. The optimal power allocation is calculated as:

$$\boldsymbol{m}_{u,d} = \frac{\boldsymbol{c}_{u,d}}{||\boldsymbol{c}_{u,d}||}\sqrt{P_d}, \tag{13}$$

Due to the asymptotic orthogonality assumption, only the allocated RISs contribute to the effective channel.

Furthermore, the coherent RIS-assisted signal power is given by

$$p_{c,u} = \left(\sum_{r \in \mathcal{R}^{(d)}} s_{u,r,d} \right)^2. \tag{14}$$

Regarding the non-coherent power we use similar steps, which results in the incoherent RIS-assisted signal power, given by:

$$p_{i,u} = \sum_{r \in \mathcal{R}^{(d)}} \gamma_{u,r}^2 \gamma_{r,d}^2 \bar{k}_{u,r}^2 M_{\mathrm{BS}} M_{\mathrm{RIS}} P_{u,r,d}, \tag{15}$$

where $\bar{k}_{u,r} = \sqrt{1/(1+K_{u,r})}$.

C. Interference Power Estimation

The interference power (analogously to the signal power) consists of two components: direct interference from other base stations and RIS-assisted interference caused by RISs not assigned to the serving base station. From the point-of-view of base station d, we consider the beamformers applied at the other interfering base stations as isotropically distributed. This makes sense, as the beamformers applied by the interfering base stations (which point to their assigned RISs) are not correlated with the interference channels.

The direct interference power is given by

$$i_{d,u} = \mathbb{E}\left[\left\| \boldsymbol{h}_{u,b}^T \boldsymbol{f}_{j_b,b} \right\|^2 \right] = \sum_{b \neq d} \gamma_{u,b}^2 P_b. \tag{16}$$

Under these assumptions, the RIS-assisted interference power is expressed as

$$i_{i,u} = \sum_{b \neq d} \sum_{r \notin \mathcal{R}^{(d)}} \gamma_{u,r}^2 \gamma_{b,r}^2 P_b M_{\mathrm{RIS}}. \tag{17}$$

Because of the isotropic beamforming assumption we do not apply averaging over the users as in (9).

The estimated achievable rate of user u served by base station d then follows as

$$\hat{r}_u^{(d)} = \log_2\left(1 + \hat{\mathrm{SINR}}_u^{(d)}\right). \tag{18}$$

D. Utility Function

The utility function used for RIS allocation is based on the mean achievable rate of the users served by each base station. For base station b, the utility associated with a given RIS allocation $\mathcal{R}^{(b)}$ is defined as

$$\mathrm{Util}^{(b)}(\mathcal{R}^{(b)}) = \frac{1}{|\mathcal{U}^{(b)}|} \sum_{u \in \mathcal{U}^{(b)}} \hat{r}_u^{(b)}. \tag{19}$$

This utility captures the average service quality experienced by the users associated with base station b.

E. Auction Format

RIS allocation among base stations is performed using a simultaneously ascending auction, which provides a low-complexity alternative to combinatorial mechanisms such as Vickrey–Clarke–Groves [10] while capturing competitive interactions. A similar RIS auction framework was considered in [13].

The auction proceeds in discrete rounds t. In each round, the auctioneer announces a uniform price p_t, increased by a fixed increment Δ_p from the previous round. Each base station b submits a binary bid vector $\boldsymbol{b}_t^{(b)} \in \{0,1\}^{N_{\mathrm{RIS}}}$, where $\boldsymbol{b}_t^{(b)}[r] = 1$ indicates willingness to bid for RIS r at price p_t.

RISs receiving a single bid are allocated at the current price; RISs with multiple bids remain contested and advance to the next round, while RISs receiving no bids remain unassigned and apply random phase shifts. An activity rule prevents strategic re-entry, i.e., a base station cannot bid for a RIS in round t if it did not bid for it in round $t-1$ [18]. The activity rule supports the identification of preferences among agents. The auction terminates once no RIS receives multiple bids.

IV. Bidding Strategies

Let $\mathcal{R}_{t-1}^{(b)}$ denote the set of RISs already allocated to base station b in previous rounds. The set of RISs that remain available at round t is given by

$$\mathcal{R}_t = \{1, \ldots, N_{\mathrm{RIS}}\} \setminus \bigcup_b \mathcal{R}_{t-1}^{(b)}. \tag{20}$$

Ideally, a base station would evaluate the utility of all possible subsets of remaining RISs; however, this combinatorial evaluation becomes infeasible as the number of RISs grows. We therefore adopt a simplified marginal approach, in which each base station estimates the utility gain of acquiring a single additional RIS, assuming no other RIS is acquired in the same round.

A. Marginal Utility Value Estimation

The estimated marginal utility value of RIS $r \in \mathcal{R}_t$ for base station b at auction round t is defined as

$$V_t^{(b)}(r) = \mathrm{Util}^{(b)}\left(\mathcal{R}_{t-1}^{(b)} \cup \{r\}\right) - \mathrm{Util}^{(b)}\left(\mathcal{R}_{t-1}^{(b)}\right). \tag{21}$$

This value represents the expected improvement in the mean achievable rate of the users served by base station b if RIS r were to be allocated to it.

B. Normalization and Standardization

While marginal utility values capture the relative desirability of RISs, their absolute scale depends on the channel realization, user distribution, and current allocation. Consequently, the magnitude of $V_t^{(b)}(r)$ can vary significantly across RISs, auction rounds, and training episodes. To obtain a bounded and numerically stable representation suitable for learning-based bidding, we apply a two-step standardization procedure.

First, negative marginal values are clipped using a rectified linear unit (ReLU):

$$\tilde{V}_t^{(b)}(r) = \max\left(V_t^{(b)}(r),\, 0\right), \tag{22}$$

such that only RISs expected to yield a performance improvement are considered. A value of zero therefore indicates that no utility gain is anticipated from acquiring RIS r in the current round.

Second, the remaining values are normalized by the maximum positive marginal gain among all available RISs,

$$V_t^{(b)}(r) \leftarrow \begin{cases} \dfrac{\tilde{V}_t^{(b)}(r)}{\max\limits_{r' \in \mathcal{R}_t} \tilde{V}_t^{(b)}(r')}, & \text{if } \max\limits_{r' \in \mathcal{R}_t} \tilde{V}_t^{(b)}(r') \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

This normalization maps marginal utility values to the interval $[0, 1]$, where one corresponds to the RIS with the highest expected utility gain in the current auction round. The relative ranking of RISs is preserved, while numerical consistency across environments and training episodes is ensured.

C. RL-Based Bidding

To enable adaptive and fairness-aware bidding, we model the auction as a multi-agent reinforcement learning problem in which each base station acts as an autonomous agent. Through repeated interaction with the auction environment, agents learn bidding strategies that account for both their own utility gains and the relative performance of other base stations.

Unlike purely local strategies, this formulation enables implicit coordination via shared information provided by the auctioneer. In particular, agents are informed of their relative service quality through a fairness-aware weighting mechanism that biases bidding toward weaker-performing cells.

1) States: The complete environment state at auction round t is defined as

$$\mathcal{S}_t = \Big(p_t,\ \big\{ V_t^{(b)}(r),\ B_t^{(b)},\ w_t^{(b)} \big\}_{\forall b,\, r} \Big), \tag{24}$$

where p_t denotes the current auction price, $B_t^{(b)}$ is the remaining budget of base station b, $V_t^{(b)}(r)$ are the normalized marginal utility values, and $w_t^{(b)}$ is a fairness weight associated with base station b.

The fairness weights are computed centrally based on the current utility values of all base stations and are defined as

$$w_t^{(b)} = \frac{\Big(\mathrm{Util}^{(b)}(\mathcal{R}_{t-1}^{(b)}) \Big)^{\gamma}}{\sum\limits_{b'} \Big(\mathrm{Util}^{(b')}(\mathcal{R}_{t-1}^{(b')}) \Big)^{\gamma}} \cdot N_{\mathrm{BS}}, \tag{25}$$

where $\gamma \geq 0$ controls the strength of the fairness mechanism. For $\gamma = 0$, all base stations are assigned identical weights, whereas larger values of γ increasingly emphasize performance differences between base stations. The normalization ensures a unit average fairness weight across base stations, aiming to stabilize the total price expenditure without strictly enforcing constancy.

2) Observations: Each agent operates on an individual observation derived from the global state. The observation available to base station b at round t is given by

$$\mathcal{O}_t^{(b)} = \Big(p_t,\ B_t^{(b)},\ w_t^{(b)},\ \big\{ V_t^{(b)}(r)\ \big\}_{\forall r} \Big). \tag{26}$$

The observation includes the fairness weights, enabling agents to condition their bidding behavior on the relative performance of other cells. This information exchange is mediated by the auctioneer and does not require direct communication between base stations.

To ensure a fixed-length observation vector, the marginal values $V_t^{(b)}(r)$ are set to -1 for RISs that are no longer available or for which base station b is inactive due to the enforced activity rule.

3) Actions: At each auction round, each agent selects a binary bid vector

$$\boldsymbol{b}_t^{(b)} \in \{0, 1\}^{N_{\mathrm{RIS}}}, \tag{27}$$

where $\boldsymbol{b}_t^{(b)}[r] = 1$ indicates that base station b places a bid for RIS r at the current price p_t, and $\boldsymbol{b}_t^{(b)}[r] = 0$ otherwise.

4) Reward Function: The reward function is designed to encourage agents to bid for RISs that provide high expected utility gains, while discouraging excessive or budget-violating bidding behavior. For base station b at auction round t, the reward is defined as

$$r_t^{(b)} = R_{1,t}^{(b)} - \beta w_t^{(b)} \big(R_{2,t}^{(b)} + R_{3,t}^{(b)} \big), \tag{28}$$

where β is a constant that controls the overall aggressiveness of bidding [14]. Rewards are evaluated before the auction outcome, i.e., after each bidding decision instead of only upon winning a RIS, enabling dense and immediate feedback during training. The three reward components are specified as follows.

The first component captures the total expected value of the bids placed by the agent in the current round:

$$R_{1,t}^{(b)} = \sum_{r=1}^{N_{\mathrm{RIS}}} V_t^{(b)}(r)\, \boldsymbol{b}_t^{(b)}[r]. \tag{29}$$

This term rewards the agent for selecting RISs that are expected to improve its utility.

The second component penalizes the monetary cost of the bids placed in the current round:

$$R_{2,t}^{(b)} = p_t \sum_{r=1}^{N_{\mathrm{RIS}}} \boldsymbol{b}_t^{(b)}[r]. \tag{30}$$

This term discourages agents from placing unnecessary bids and promotes selective bidding behavior.

The third component introduces an additional penalty if the total bid cost exceeds the remaining budget of the base station:

$$R_{3,t}^{(b)} = 2 \cdot \max\left(p_t \sum_{r=1}^{N_{\mathrm{RIS}}} \boldsymbol{b}_t^{(b)}[r] - B_t^{(b)},\ 0 \right). \tag{31}$$

This term explicitly discourages budget violations by penalizing bids that exceed the available budget. The scaling factor emphasizes budget violations relative to regular bidding costs.

Scaling the cost-related terms $R_{2,t}^{(b)}$ and $R_{3,t}^{(b)}$ by the fairness weight $w_t^{(b)}$ biases bidding toward weaker-performing base stations, penalizing aggressive bids from stronger agents while allowing weaker agents to bid more aggressively. This

promotes a more balanced RIS allocation while preserving competition.

The proposed fairness mechanism assumes truthful reporting of performance-related information to the auctioneer. Strategic misreporting could introduce vulnerabilities, as a base station claiming lower performance would reduce its own fairness weight and make competing agents more conservative, potentially gaining an advantage. Addressing such behavior would require additional mechanism-design measures, such as verification or incentive-compatible reporting, which are beyond the scope of this work.

D. Implementation

The environment is implemented using the Gymnasium interface [19] with multi-agent support provided by PettingZoo [20] and SuperSuit [21]. Each training episode corresponds to a complete auction process, starting from the initial price and ending when RISs receive no further bids.

Agents are trained using a policy-gradient-based method, specifically the Proximal Policy Optimization (PPO) algorithm [22] as implemented in Stable-Baselines3 [23]. An undiscounted return is employed, reflecting the finite-horizon nature of the auction process, where all decisions within an episode contribute equally to the final allocation outcome. Aside from the undiscounted return, all other hyperparameters were kept at the default values of the PPO implementation. Due to the nature of the PPO implementation CPU-based training was utilized to ensure optimal execution speed.

Training is performed episodically across a large number of independent network realizations, where user locations, RIS positions, and channel parameters are randomized between episodes. The implementation supports a variable number of base stations, users, and RISs.

The implementation will be made available upon acceptance of the work at: https://github.com/MartinMarkZan.

V. Simulations

In this section, we evaluate the performance of the proposed fairness-aware RIS allocation framework through numerical simulations. The focus is on quantifying the trade-off between system efficiency and user fairness, as well as on illustrating how the proposed mechanism redistributes RIS resources across base stations.

A. Simulation Setup

We consider a two-base-station scenario with a fixed number of users and RIS elements. The base stations are located at the opposite edges of the region of interest, while the RISs are deployed along the cell edge on a straight line. The placement of the RISs increases the competition between the two base stations. User locations are generated uniformly within the cell area. One of the base stations is overloaded with users (denoted as BS0 in the figures); on average, it serves approximately three times as many users as the other base station (BS1). While this initial study focuses on a two-base-station scenario with RISs positioned at the cell edge to clearly illustrate fundamental fairness-performance trade-offs, future work will involve more complex network topologies with a larger number of base stations, users, and diverse RIS placements.

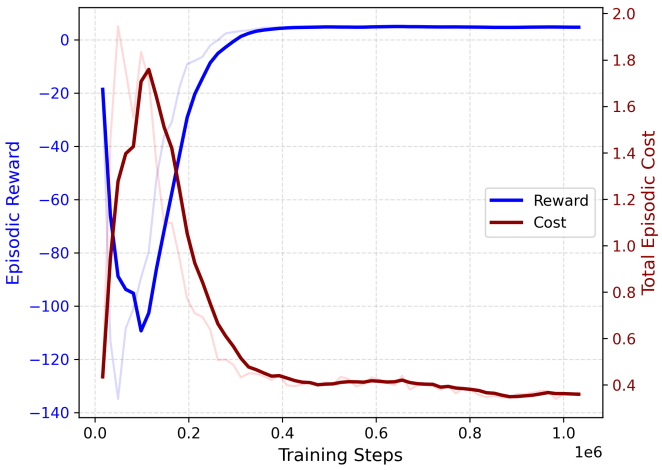

Fig. 1. Convergence of the episodic reward (left axis) and total auction cost (right axis) during training. Solid lines represent moving-average smoothed curves (window size=5), while semi-transparent lines show the raw data.

TABLE I
Simulation Parameters

Carrier frequency	$f_c = 26\,\text{GHz}$
Number of base stations	$N_{\text{BS}} = 2$
Number of base station antennas	$M_{\text{BS}} = 50$
Number of users	$N_{\text{UE}} = 20$
Number of RISs	$N_{\text{RIS}} = 10$
Number of RIS elements	$M_{\text{RIS}} = 250$
Transmit power per subcarrier	$P = 100\,\text{mW}$
Subcarrier bandwidth	$15\,\text{kHz}$
AWGN noise power spectral density	$-174\,\text{dBm/Hz}$
Noise figure	$6\,\text{dB}$
Path-loss exponent under LOS (NLOS)	2 (4.5)
K-factor under LOS (NLOS)	100 (3)
Distance-dependent LOS-probability	$p_{\text{LOS}}(d) = e^{-d/25}$
Shadow fading variance	$10\,\text{dB}$
Auction initial price	$p_0 = 0.05$
Auction price increment	$\Delta p = 0.05$
Budget	$B_0^{(b)} = 1$

As illustrated in Fig. 1, the reinforcement learning agents converged to stable reward values during training, indicating reproducible learning behavior. The curves show that agents reliably discover effective bidding policies that maximize gains while maintaining stable budget utilization, confirming robustness. During evaluation, we used 200 macroscopic realizations (user positions and large-scale path gains) and 20 independent microscopic fading realizations per macroscopic setup to aggregate the results.

Table I summarizes the main simulation parameters. The geometry is shown in Fig. 2 for a random snapshot of user positions.

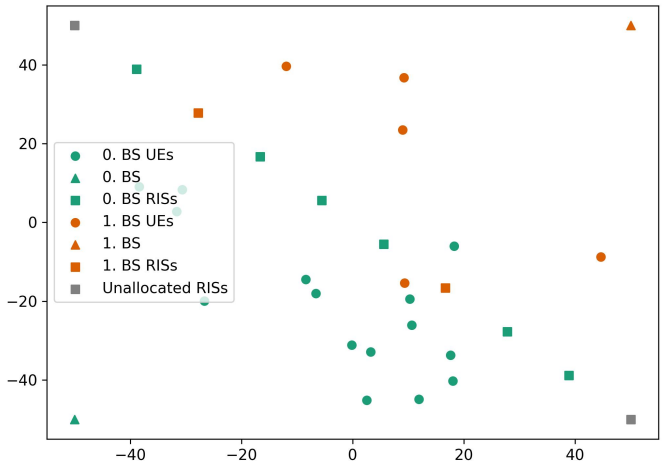

Fig. 2. Representative network realization for $\gamma = 0.2$, showing the locations of the two base stations, users, allocated RISs, and unassigned RISs.

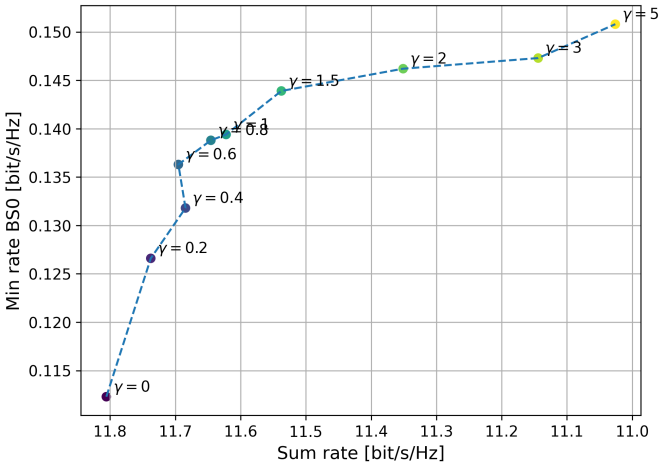

Fig. 3. Trade-off between sum rate and the minimum user rate of the overloaded base station (BS0). Each point corresponds to a model with a different value of the fairness strength γ.

B. Efficiency-Fairness Trade-off

Fig. 3 visualizes the trade-off between system efficiency and fairness using the sum rate versus the minimum user rate of the overloaded base station.

As γ increases, the operating point moves along a Pareto-like frontier: the minimum rate of BS0 improves by approximately 34%, while the sum rate of the two base stations decreases only moderately (less than 7% over the considered range). This confirms that the proposed mechanism is able to substantially improve the performance of the worst-served users without causing a severe loss in overall system throughput.

C. Fairness Evaluation via Atkinson Index

To quantify fairness more systematically, Fig. 4 reports the Atkinson inequality index as a function of the fairness strength

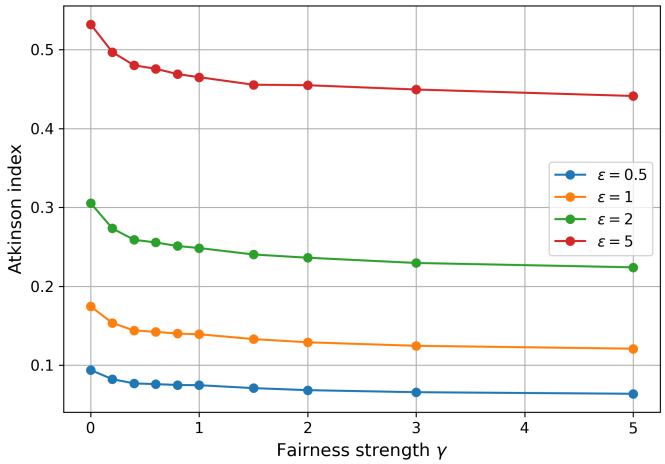

Fig. 4. Atkinson inequality index as a function of the fairness strength γ for different values of the sensitivity parameter ϵ, which controls the emphasis on low-rate users.

γ. The index is defined as

$$A_\varepsilon(y_1, \ldots, y_N) = 1 - \frac{E_\varepsilon}{\mu}, \tag{32}$$

where

$$E_\varepsilon = \begin{cases} \left(\frac{1}{N}\sum_{i=1}^{N} y_i^{1-\varepsilon}\right)^{\frac{1}{1-\varepsilon}}, & 0 \le \varepsilon \neq 1, \\ \left(\prod_{i=1}^{N} y_i\right)^{\frac{1}{N}}, & \varepsilon = 1, \\ \min(y_1, \ldots, y_N), & \varepsilon = +\infty, \end{cases}$$

and μ is the mean of the input values. The Atkinson index takes values in $[0, 1]$, where smaller values indicate more equal rate distributions.

For all considered ϵ, increasing γ consistently reduces the inequality index, confirming that the proposed framework improves fairness across users. Larger values of ϵ result in higher Atkinson indices, since the metric places greater emphasis on the worst-served users and penalizes residual disparities more strongly. The monotonic decrease of all curves with γ demonstrates that the fairness improvement is robust with respect to the chosen fairness sensitivity.

D. RIS Allocation Behavior

Fig. 5 shows the average RIS allocation as a function of the fairness parameter γ. The figure reports the number of RISs assigned to BS0, BS1, and those remaining unallocated.

As γ increases, RIS resources are progressively shifted from BS1 to the overloaded BS0, which directly explains the observed improvement in the minimum rate of BS0. At the same time, the number of unallocated RISs decreases, indicating more aggressive bidding behavior by the weaker-performing base station and a less competitive auction.

For the considered operating point, the total price spent by the two base stations remains approximately constant across γ. However, this behavior is not guaranteed in general. Additional

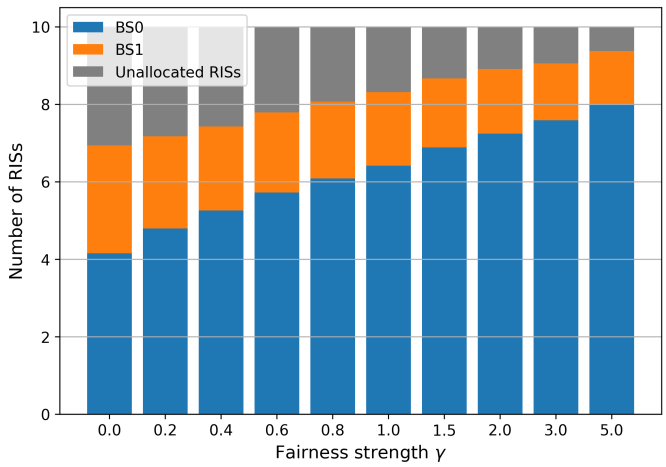

Fig. 5. Increasing γ shifts RIS resources from BS1 to the overloaded BS0, while also decreasing the number of unallocated RISs due to more aggressive bidding behavior.

experiments with different cost-scaling parameters (β) show that stronger fairness pressure can also lead to increased overall expenditure. This highlights an inherent interaction between fairness objectives and economic efficiency, which can be controlled through appropriate reward design.

VI. Conclusion

In this work, we investigated auction-based allocation of reconfigurable intelligent surfaces in asymmetric multi-cell networks with a focus on fairness-aware resource distribution. RISs were modeled as shared infrastructure and dynamically allocated through a simultaneous ascending auction, enabling scalable allocation in competitive cell-edge scenarios. To address performance imbalances caused by uneven user distributions, we proposed a cooperative multi-agent reinforcement learning framework that integrates a performance-dependent fairness mechanism into the bidding process.

Simulation results show that the proposed approach significantly improves the performance of the worst-served users, while maintaining competitive sum-rate performance. By adjusting the fairness parameter, the trade-off between data rate and equitable resource allocation can be explicitly controlled. These findings demonstrate the effectiveness of combining auction-based mechanisms with cooperative reinforcement learning for fair and efficient RIS utilization in future wireless networks.

While the proposed framework demonstrates strong performance in moderate-sized scenarios, extending the approach to large-scale networks with a higher number of base stations, users, and RISs remains an important research direction. Additionally, other auction formats, such as sealed-bid mechanisms or dynamic pricing schemes, could be investigated to further improve efficiency or fairness under different deployment assumptions. Future work can also consider non-stationary environments with time-varying users.

References

[1] F. Irram, M. Ali, Z. Maqbool, F. Qamar and J. J. Rodrigues, "Coordinated Multi-Point Transmission in 5G and Beyond Heterogeneous Networks," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-6, doi: 10.1109/INMIC50486.2020.9318091

[2] S. Schwarz and M. Rupp, "Exploring Coordinated Multipoint Beamforming Strategies for 5G Cellular," in IEEE Access, vol. 2, pp. 930-946, 2014, doi: 10.1109/ACCESS.2014.2353137

[3] E. Nayebi, A. Ashikhmin, T. L. Marzetta and H. Yang, "Cell-Free Massive MIMO systems," 2015 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2015, pp. 695-699, doi: 10.1109/ACSSC.2015.7421222

[4] C. F. Mendoza, M. Kaneko, M. Rupp and S. Schwarz, "Enhancing the Uplink of Cell-Free Massive MIMO through Prioritized Sampling and Personalized Federated Deep Reinforcement Learning," in IEEE Transactions on Cognitive Communications and Networking, vol. 12, pp. 395-411, 2026, doi: 10.1109/TCCN.2025.3561289

[5] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau and K. V. Srinivas, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions," in IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 611-652, Firstquarter 2022, doi: 10.1109/COMST.2021.3135119

[6] C. F. Mendoza, S. Schwarz and M. Rupp, "User-Centric Clustering in Cell-Free MIMO Networks using Deep Reinforcement Learning," ICC 2023 - IEEE International Conference on Communications, Rome, Italy, 2023, pp. 1036-1041, doi: 10.1109/ICC45041.2023.10279626

[7] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," IEEE Commun. Lett., vol. 25, no. 1, pp. 269–273, Jan. 2021.

[8] M. A. Msleh, F. Heliot, and R. Tafazolli, "Ergodic capacity analysis of reconfigurable intelligent surface assisted MIMO systems over Rayleigh-Rician channels," IEEE Commun. Lett., vol. 27, no. 1, pp. 75–79, Jan. 2023.

[9] Q. N. Le, V.-D. Nguyen, O. A. Dobre, and R. Zhao, "Energy efficiency maximization in RIS-aided cell-free network with limited backhaul," IEEE Commun. Lett., vol. 25, no. 6, pp. 1974–1978, Jun. 2021.

[10] N. Nisan and A. Ronen, "Algorithmic mechanism design," Games Econ. Behav., vol. 35, no. 1, pp. 166–196, 2001.

[11] P. Cramton and A. Ockenfels, "The German 4G Spectrum Auction," The Economic Journal, 127 (October), F305–F324, 2017.

[12] P. Milgrom, "Putting Auction Theory to Work", Stanford University, California, Cambridge University Press, 2004.

[13] S. Schwarz, "Gambling on Reconfigurable Intelligent Surfaces," IEEE Communications Letters, vol. 28, no. 4, pp. 957–961, April 2024.

[14] M. M. Zan and S. Schwarz, "Auction-Based RIS Allocation With DRL: Controlling the Cost-Performance Trade-Off," to be published in IEEE Open Journal of the Communications Society 2026.

[15] H. Zhang, W. Wang, H. Zhou, Z. Lu and M. Li, "A Hierarchical DRL Approach for Resource Optimization in Multi-RIS Multi-Operator Networks," 2025, arXiv:2410.12320

[16] V. Nanduri and T. K. Das, "A Reinforcement Learning Model to Assess Market Power Under Auction-Based Energy Pricing," in IEEE Transactions on Power Systems, vol. 22, no. 1, pp. 85-95, Feb. 2007

[17] C. A. Balanis, "Antenna Theory", Hoboken, New Jersey, U.S.: John Wiley & Sons, Inc., 2005.

[18] T. Roughgarden, Twenty Lectures Algorithmic Game Theory. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[19] M. Towers et al. (Mar. 2025). Gymnasium. [Online]. Available: https://zenodo.org/records/14983111

[20] J. K. Terry et al., "PettingZoo: Gym for multi-agent reinforcement learning," 2021, arXiv:2009.14471.

[21] J. K. Terry, B. Black, and A. Hari, "SuperSuit: Simple microwrappers for reinforcement learning environments," 2020, arXiv:2008.08932.

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, arXiv:1707.06347.

[23] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," J. Mach. Learn. Res., vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html