Dynamic Energy Efficient Resource Allocation for Massive MIMO Networks Using Randomized Ensembled Double Q-Learning Algorithm

Zhikai Liu, Navneet Garg, *Senior Member, IEEE*, and Tharmalingam Ratnarajah, *Senior Member, IEEE*

***Abstract*—This paper tackles the challenge of power consumption in the massive multiple-input multiple-output (mMIMO) base station (BS), where continuous operation of all antennas generates significant heat within a limited physical area. We propose a strategic power control scheme to enhance energy efficiency and mitigate thermal impact. Our approach introduces a time-slotted model incorporating dynamic, time-varying user quality of service (QoS) requirements. We examine energy efficiency under various conditions, presenting discrete and analog power allocation methods for both hybrid and fully digital precoding, with consideration of hardware impairments (HWI). We frame the energy efficiency optimization as dynamic Markov decision process (MDP) problems, constrained by total power, per-antenna power, and dynamic QoS requirements. The randomized ensembled double *Q*-learning (REDQ) algorithm is utilized with an action coding scheme to reduce computational complexity. By comparing existing reinforcement learning algorithms and evaluating our proposed power allocation schemes across diverse scenarios, simulations demonstrate that our approach improves energy efficiency effectively under varying operational conditions, showcasing its potential as a robust solution for adaptive resource allocation in mMIMO systems.**

***Index Terms*—Dynamic power allocation, energy efficiency, fully digital precoding, hybrid precoding, hardware impairment, reinforcement learning.**

I. Introduction

THE MMIMO systems facilitate the deployment of hundreds of antennas at a single BS, significantly enhancing spectral efficiency [1], [2]. However, achieving high energy efficiency in mMIMO systems presents significant challenges due to high power consumption at individual elements and across the entire array [3], [4], [5]. This issue is exacerbated by substantial fluctuations in user traffic load throughout the day [6], indicating that maintaining maximum power allocation to each antenna may not be necessary at all times. To address this, implementing a dynamic power allocation scheme, tailored to users' changing demands, is essential. It should be noted that in cognitive radio, advanced learning, and decision-making capabilities play a crucial role in dynamic spectrum management and optimizing power allocation to enhance communication efficiency [7]. Dynamic power allocation schemes enable mMIMO systems to adjust their power distribution in real time based on varying user demands and traffic patterns throughout the day. By monitoring and predicting these temporal changes, the system can selectively allocate power to different antennas, reducing power consumption when traffic is low and boosting it when demand increases. This approach not only enhances energy efficiency but also ensures that the system maintains excellent performance and QoS by dynamically balancing the load and mitigating potential interferences. Consequently, dynamic power allocation helps in minimizing thermal impact and prolonging the lifespan of the hardware components by preventing overheating, while still meeting the varying QoS requirements of different users efficiently.

Specifically, power allocation schemes can be categorized into two types: analog and discrete. Analog power allocation involves real-valued power levels that can be finely adjusted to accommodate system requirements [8]. Its flexibility allows the system to respond accurately and swiftly to changes in channel conditions, optimizing system capacity and link reliability. However, the requirement for a complex, high-precision variable gain amplifier can be costly and may increase the system's overall power consumption. Furthermore, operating all antennas simultaneously leads to thermal exhaustion, potentially shortening the system's lifespan. Conversely, discrete power allocation involves a finite set of power levels, making it simpler to manage and implement with less expensive hardware [9]. This method allows for turning off some antennas, mitigating heat generation and thermal exhaustion. The approach significantly reduces the complexity and cost of the power amplifier design, which in turn lowers power consumption and boosts energy efficiency. Therefore, comparing and analyzing analog and discrete power allocation is crucial for mMIMO systems. Such analysis can highlight promising directions for future investigation and

Received 20 February 2024; revised 29 July 2024 and 21 November 2024; accepted 28 December 2024. Date of publication 1 January 2025; date of current version 8 August 2025. The associate editor coordinating the review of this article and approving it for publication was G. Ding. *(Corresponding author: Zhikai Liu.)*

Zhikai Liu is with the Institute for Data, Image, and Communications, School of Engineering, The University of Edinburgh, EH8 9YL Edinburgh, U.K. (e-mail: zliu-129@sms.ed.ac.uk).

Navneet Garg was with the Institute for Data, Image, and Communications, School of Engineering, The University of Edinburgh, EH8 9YL Edinburgh, U.K. He is now with the Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur 302004, India (e-mail: navneet.garg4@gmail.com).

Tharmalingam Ratnarajah was with the Institute for Data, Image, and Communications, School of Engineering, The University of Edinburgh, EH8 9YL Edinburgh, U.K. He is now with the Electrical and Computer Engineering, College of Engineering, San Diego State University, San Diego, CA 92182 USA (e-mail: t.ratnarajah@ieee.org).

Digital Object Identifier 10.1109/TCCN.2024.3524640

development, paving the way for more efficient and sustainable operations.

Another critical aspect of improving energy efficiency in mMIMO systems is the precoding method, which can be divided into fully digital and hybrid precoding. Fully digital precoding employs baseband (BB) digital signal processing to generate a set of precoded signals for transmission. Research such as [10] proposed an energy efficiency maximization-based joint antenna and user selection algorithm for Multi-user (MU) massive-MIMO downlink with digital zero-forcing (ZF) precoding. In a related study, the authors in [11] tackled the NP-hard discrete optimization problem of performing transmit antenna selection in a single-cell MU mMIMO system's downlink. The work in [12] devised an algorithm that maximizes the energy efficiency of digital precoding using the dual Lagrangian method. Similarly, a joint antenna selection and power allocation for fully digital precoding was proposed in [13]. However, these studies consider a relatively small number of antennas, and digital precoding for mMIMO systems with a large number of antennas may result in high complexity and power consumption. Hybrid precoding, on the other hand, offers a viable alternative, blending the benefits of fully digital and analog precoding. Dividing the precoding operation into a digital BB part and an analog RF part can reduce hardware complexity and power consumption. For energy efficiency, [14] proposed a method to maximize the non-concave global energy efficiency in the mmWave mMIMO system using fractional programming. Other notable contributions include a novel swarm intelligence-based power allocation technique for the MU-mMIMO system [15], a deep learning-based power allocation and hybrid precoding algorithm via a fully connected deep neural network (DNN) [16], and an investigation into ZF hybrid precoding to maximize the downlink sum-rate based on the Karush-Kuhn-Tucker (KKT) condition [17]. Lastly, [18] presented a power allocation scheme for hybrid precoding to reduce the bit error rate (BER).

Despite the considerable contributions in the above literature, few studies systematically compare the energy efficiency in the presence and absence of HWI. These imperfections in HWI, which include phase noise, in-phase/quadrature (I/Q) imbalance, and power amplifier nonlinearities, can profoundly influence the system energy efficiency [19]. Consequently, it is imperative to incorporate HWI when designing power allocation schemes for mMIMO systems.

In summary, the research on energy efficiency in mMIMO systems should address dynamic user traffic load requirements, explore both analog and discrete power allocation, compare fully digital and hybrid precoding methods, and consider the impact of HWI. This paper aims to fill these gaps by presenting a comprehensive study on these aspects, offering insights into their operational nuances, performance benchmarks, and cost-efficiency tradeoffs.

Remark (Comparison of Fully Digital and Hybrid Precoding Technologies): There are fundamental differences between fully digital and hybrid precoding architectures. Fully digital precoding uses an independent RF chain for each antenna element, enabling precise control and excellent performance in signal processing and beam shaping [20], [21]. In contrast, hybrid precoding employs fewer RF chains, each controlling a subarray of antenna elements through phase shifters [22], trading off flexibility for reduced hardware complexity and cost. Our work compares these technologies to offer a comprehensive perspective on their advantages and limitations in different operational contexts. Fully digital precoding is ideal for scenarios demanding high precision and adaptability, while hybrid precoding is better suited for cost-sensitive deployments with more straightforward hardware requirements. By juxtaposing these approaches, we aim to provide valuable insights into their operational nuances, performance benchmarks, and cost-efficiency tradeoffs, contributing to advancing precoding techniques.

A. Contributions

In this paper, we formulate and solve the energy efficiency maximization problem in mMIMO networks under time-varying dynamic user QoS requirements. We present a multi-user mMIMO system model utilizing both fully digital and hybrid precoding techniques considering the presence and absence of HWI. Under a time-slotted model, the long-term energy efficiency objective is expressed as a dynamic MDP problem with constraints including total power, per-antenna power, and the user's QoS requirements in terms of Signal-to-Interference-plus-Noise Ratio (SINR). To solve this constrained optimization problem for both analog and discrete power allocation, the REDQ algorithm is utilized with a proposed action coding scheme. Through extensive simulations, our proposed power allocation scheme based on the REDQ algorithm can achieve more efficient power allocation, outperforming baseline reinforcement learning algorithms including SAC (Soft Actor-Critic) [23], TD3 (Twin Delayed Deep Deterministic Policy Gradient) [24], and DDPG (Deep Deterministic Policy Gradient) [25]. These simulation results provided a comprehensive comparison of various scenarios under multiple constraints.

- We consider a time-slotted model with varying channel conditions and dynamic users' QoS requirements. By incorporating both fully digital and hybrid precoding in the system model, we address the practical challenge of overheating in large array systems through discrete and analog power allocation strategies. The impact of HWI on energy efficiency is also considered, making our approach relevant for real-world deployments. The inclusion of per-antenna power constraints, along with total transmit power constraints, ensures that antennas do not use excessive power under certain channel conditions, enhancing the reliability and longevity of the hardware.
- We formulate the optimization as a long-term energy efficiency problem, considering channel state variations as an MDP with unknown transition probabilities. To solve this, we propose an action coding scheme in conjunction with the REDQ algorithm to obtain discrete and analog power allocation solutions. This action coding scheme addresses the challenge of exponentially large action spaces resulting from numerous antennas and discrete power levels per antenna. Additionally, the

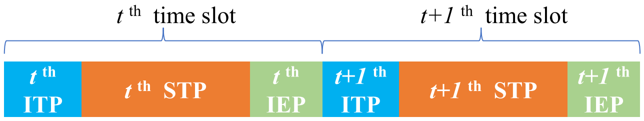

Fig. 1. Time slotted model.

minimum transmit power is explored to further reduce the action space and meet QoS constraints, ensuring that our solutions are scalable and practical for large-scale mMIMO systems.

- We conduct extensive simulations to validate the proposed analog and discrete power control solutions for both fully digital and hybrid precoding with and without HWI. The results show that the obtained power allocation schemes satisfy the transmit power constraint, providing a feasible solution for enhancing energy efficiency in dynamic mMIMO systems. The complexity experiments present the significance of the action coding method, especially for scenarios with a large number of antennas.

B. Organization and Notations

The remainder of this paper unfolds as follows: Section II introduces the system model, encapsulating the time-slotted model and the research scenario. Section III delves into power allocation analysis for fully digital precoding, while Section IV does the same for hybrid precoding. Section V presents the formulation of the problem. The reinforcement learning solution is delineated in Section VI, and Section VII discusses the results derived from our simulations. Finally, Section VIII concludes the paper, summarizing key findings and insights.

In this paper, scalars, vectors, matrices, and sets are represented by the lower case (a), lower case boldface ($\mathbf{a}$), upper case boldface ($\boldsymbol{A}$), and calligraphic ($\mathcal{A}$) letters respectively. Transpose of matrices is denoted by $(.)^T$. The notation $\|.\|_2$ denotes the l_2 norm. $|\mathcal{K}|$ denotes the cardinality of the set $\mathcal{K}$. The main variables in this paper are listed in Table I.

II. System Model

We consider a single BS equipped with M antennas servicing K downlink users, each with N antennas to decode N_d data stream. Let $\mathcal{M}$ and $\mathcal{K}$ represent the set of transmitting antennas and users.

Our proposed system utilizes a time-slotted dynamic model as Figure 1, to outline the structure of each time interval. We consider time-varying channels across time slots as [26], [27]. Within a time slot, the channel remains constant. Time-varying channels are modeled using finite-state Markov chains, where the ergodic channel in each time slot takes values in one of the Markov states. At the onset of each time slot, the information detected phase (IDP) is carried out, where the BS detects the number of active users and evaluates their respective load demands within its coverage area. The traffic load demand is represented as the QoS constraint, formulated as SINR constraint $\bar{\xi}$ in the following sections, wherein the SINR of each user must exceed their respective SINR constraint. Importantly, these QoS constraints dynamically fluctuate in different time slots. The QoS constraint value can be 0, signifying that the user is inactive during this time slot. After analyzing QoS requirements, during the signal transmitted phase (STP), the BS implements the scheduling algorithm to apportion the available resources among the users and then initiates signal transmission to the users. Upon receipt of these signals, in the information exchange phase (IEP), the users reciprocate by transmitting feedback information, encompassing aspects such as signal quality and power, back to the BS. This feedback allows the BS to update the beamforming vectors and power settings accordingly. Leveraging the feedback and historical data, the system utilizes a learning algorithm to enhance decision-making in future time slots.

We assume the channel state information (CSI) variations follow a Markov process. At the beginning of each time slot, the BS performs channel estimation to obtain the CSI. This estimation process includes pilot signal transmission and user feedback during the IEP. The estimated channel is obtained using least squares. The CSI obtained is used to update the beamforming vectors and adjust power settings. For beamforming and signal transmission, we employ a uniform planar array (UPA) at the BS.

TABLE I
List of Main Variables

Symbol	Description
M	Antenna number of BSs
N	Antenna number of user devices
ξ	SINR constraint
K	User number
N_d	Data stream number
$\mathbf{x}$	Transmitted signal
$\boldsymbol{E}$	Amplifier Non-linearity and Mutual Coupling Matrix
$\boldsymbol{V}$	Precoding Matrix
$\mathbf{s}$	Data Symbols Vector
$\mathbf{c}$	HWI Noise Vector
P_0	Power upper limit at the BS for fully digital precoding
P_0'	Power upper limit at the BS for hybrid precoding
$\boldsymbol{P}$	Power allocation matrix for fully digital precoding
$\boldsymbol{P}'$	Power allocation matrix for hybrid precoding
μ	Path Loss Exponent for Mutual Coupling
δ	Quantization Step Size
C_{tot}	Total transmitted power for fully digital precoding
C_{tot}'	Total transmitted power for hybrid precoding
$\mathbf{y}_k$	Received signal of the user k
η	Energy efficiency for fully digital precoding
η'	Energy efficiency for hybrid precoding
$\boldsymbol{H}$	Channel matrix for fully digital precoding
$\boldsymbol{H}'$	Channel matrix for hybrid precoding
$\mathbf{n}_k$	Additional Gaussian noise for user k
T_s	Duration of an OFDM Symbol
$\boldsymbol{F}_{\text{RF}}$	RF Precoder Matrix
$\mathbf{f}_{\text{BB}}$	BB Precoding Vector
$\mathbf{w}_{\text{RF}}$	RF Combiner
$\boldsymbol{F}_{\text{BB}}$	MMSE BB Precoder Matrix
γ	Discount factor
$\mathbf{a}, \mathbf{s}, r$	Action, state and reward
J	Number of Q-functions
J'	The subset size of updated Q-functions
D	UTD ratio
ω	Learning rate
ρ	Rate of target parameter updates
α	Entropy term scaling for exploration
$\mathcal{B}$	Mini batch set
$\mathcal{O}$	Experience buffer
N_e	Number of training steps

III. Fully Digital Precoding Model

A. Transmitted Signal

In the fully digital model, the transmitted signal from the BS with transmit hardware impairments can be written as [28]

$$\mathbf{x} = \boldsymbol{E}\boldsymbol{V}\mathbf{s} + \mathbf{c}, \tag{1}$$

where the signal vector $\mathbf{s}$ is $KN_d \times 1$ vector with zero mean and covariance matrix $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \frac{P_0}{KN_d}\mathbf{I}_{KN_d}$, where P_0 denotes the upper limit of signal power of the BS; the matrix $\boldsymbol{V} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K]$ stands for the $M \times (KN_d)$ precoding matrix and can be obtained by ZF such that the norm of each column of $\boldsymbol{V}$ is unity, i.e., $\boldsymbol{V}(:,k)^H\boldsymbol{V}(:,k) = 1$; the matrix $\boldsymbol{E} \in \mathbb{C}^{M\times M}$ models the power allocated to each transmit antenna that affects the signal during transmission. We write the entries of $\boldsymbol{E} = \boldsymbol{P}^{\frac{1}{2}} + \boldsymbol{E}_{\text{OD}}$. $\boldsymbol{P}^{\frac{1}{2}}$ is a diagonal matrix representing the power allocation across antennas, where each diagonal element $\sqrt{p_m}$ corresponds to the power level for antenna m. $\boldsymbol{E}_{\text{OD}}$ denotes an off-diagonal (OD) matrix representing mutual coupling between antennas. Thus, $\boldsymbol{E}$ can be modeled as,

$$E(m,m') = \begin{cases} \sqrt{p_m}, & m = m', \\ \mathcal{CN}\Big(0, \sigma_E^2 |m-m'|^{-\mu}\Big), & m \neq m', \end{cases} \tag{2}$$

where $\boldsymbol{P} = \mathcal{D}(p_1, \ldots, p_m, \ldots, p_M), \forall m = 1, \ldots, M$ is an $M \times M$ diagonal power allocation matrix. For simplicity, the signal coming from mutual coupling is assumed to have the power of $\sigma_E^2|m-m'|^{-\mu}$, where σ_E^2 is a small variance. μ is a decay exponent, implying that the coupling effect between antennas decreases as the index difference $|m-m'|$ increases. Further, the statistics of $\boldsymbol{E}$, which defines the relationship between power allocation and mutual coupling in the MIMO system's hardware, can be given as,

$$\mathbb{E}\{E(m,m')E^*(n,n')\} = \begin{cases} \delta_{mm'}p_m + (1-\delta_{mm'})|m-m'|^{-\mu}\sigma_E^2, & m=n, m'=n', \\ 0, & o.w., \end{cases} \tag{3}$$

and $\mathbb{E}\{\boldsymbol{E}_{\text{OD}}\} = \mathbf{0}$. $\delta_{mm'}$ is the Kronecker delta function. The Kronecker delta function equals 1 if m and m' are the same, indicating a diagonal element, and equals 0 if m and m' are different, indicating an off-diagonal element. Let $\bar{\mathbf{s}} = \boldsymbol{E}\boldsymbol{V}\mathbf{s} = [\bar{s}(1), \ldots, \bar{s}(M)]^T$, with the covariance matrix as

$$\begin{aligned} \mathbb{E}\{\bar{\mathbf{s}}\bar{\mathbf{s}}^H\} &= \frac{P_0}{KN_d}\mathbb{E}\{\boldsymbol{E}\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{E}^H\} \\ &= \frac{P_0}{KN_d}\mathbb{E}\Big\{\Big(\boldsymbol{P}^{\frac{1}{2}} + \boldsymbol{E}_{\text{OD}}\Big)\boldsymbol{V}\boldsymbol{V}^H\Big(\boldsymbol{P}^{\frac{1}{2}} + \boldsymbol{E}_{\text{OD}}\Big)^H\Big\} \\ &= \frac{P_0}{KN_d}\Big[\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{P}^{\frac{1}{2}} + \boldsymbol{G}\Big], \end{aligned} \tag{4}$$

where each entry of $\boldsymbol{G}$ can be computed as

$$\begin{aligned} G(i,j) &= \mathbb{E}\Big\{\Big[\boldsymbol{E}_{\text{OD}}\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{E}_{\text{OD}}^H\Big]_{i,j}\Big\} \\ &= \mathbb{E}\sum_{m,u} E_{\text{OD}}(i,m)\Big[\boldsymbol{V}\boldsymbol{V}^H\Big]_{mu} E_{\text{OD}}^*(j,u) \\ &= \delta_{ij}\sum_{u}\Big[\boldsymbol{V}\boldsymbol{V}^H\Big]_{uu}\mathbb{E}|E_{\text{OD}}(i,u)|^2 \\ &= \delta_{ij}\sigma_E^2\sum_{u\neq i}\Big[\boldsymbol{V}\boldsymbol{V}^H\Big]_{uu}|i-u|^{-\mu}. \end{aligned} \tag{5}$$

The $M \times 1$ vector $\mathbf{c}$ of the Eq. (1) denotes the unknown part of HWI due to quantization noise from the data converters and the other random factors. We model it as a sum of uniform and Gaussian random variables as

$$\mathbf{c} \sim U\Big(-\frac{\delta}{2}, +\frac{\delta}{2}\Big) + jU\Big(-\frac{\delta}{2}, +\frac{\delta}{2}\Big) + \mathcal{CN}\Big(0, \sigma_c^2\mathbf{I}\Big), \tag{6}$$

where δ is the quantization step and depends on the number of bits $\delta = \frac{1}{2^{N_{\text{bits}}}}$ with N_{bits} being the bits used for data converters for the quantization of real and imaginary values. The variance σ_c^2 is very small and accounts for the residual HWI errors. Here, for simplicity, we have assumed Cartesian quantization; one can also perform quantization with amplitude and phase. Also, note that $\mathbf{c}$ has zero mean with the covariance matrix $\bar{\sigma}_c^2\mathbf{I} = (\frac{\delta^2}{12} + \sigma_c^2)\mathbf{I}$, and is independent of $\mathbf{s}$ and $\boldsymbol{E}$. The modeling of $\mathbf{c}$ as a combination of uniform and Gaussian random variables is practical because it accurately represents quantization noise from data converters and residual random errors, ensuring $\mathbf{c}$ has zero mean and a realistic covariance matrix. This approach captures the variance from both noise types and assumes independence from transmitted symbols and channel estimation errors, simplifying analysis and maintaining analytical tractability. The total power constraint can be given as

$$\begin{aligned} C_{\text{tot}}(\boldsymbol{P}) &= tr\big(\mathbb{E}[\mathbf{x}\mathbf{x}^H]\big) \\ &= tr\Big(\mathbb{E}\Big[(\bar{\mathbf{s}} + \mathbf{c})(\bar{\mathbf{s}} + \mathbf{c})^H\Big]\Big) \\ &= \frac{P_0}{KN_d}tr\Big(\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{P}^{\frac{1}{2}}\Big) + \frac{P_0}{KN_d}tr(\boldsymbol{G}) + \bar{\sigma}_c^2 tr(\mathbf{I}), \\ &= \frac{P_0}{KN_d}\,\text{tr}\Big(\boldsymbol{P}\boldsymbol{H}\big(\boldsymbol{H}^H\boldsymbol{P}\boldsymbol{H}\big)^{-1}\boldsymbol{H}^H\boldsymbol{P}\Big) + \frac{P_0}{KN_d}\,\text{tr}(\boldsymbol{G}) \\ &\quad + \bar{\sigma}_c^2\,\text{tr}(\mathbf{I}) \leq P_0, \end{aligned} \tag{7}$$

$\boldsymbol{H} = [\boldsymbol{H}_1, \ldots, \boldsymbol{H}_K] \in \mathbb{C}^{M\times(KN_d)}$ represents the overall channel matrix for all users. In (7), to eliminate inter-user interference with ZF transmission, we aim to find a precoder $\boldsymbol{V}$ that satisfies $\boldsymbol{H}_k^H\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}_{-k} = \mathbf{0}, \forall k$, where $\boldsymbol{H}_k \in \mathbb{C}^{M\times N_d}$ is the channel matrix for hybrid precoding, and $\boldsymbol{V}_{-k}$ represents the precoding matrix for all users except k. This condition ensures each user's signal is orthogonal to the channels of other users, effectively nullifying inter-user interference. We define the ZF precoder as $\boldsymbol{V} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{H}(\boldsymbol{H}^H\boldsymbol{P}\boldsymbol{H})^{-1}$, where $\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{H}$ incorporates per-antenna power allocation via $\boldsymbol{P}$, and $(\boldsymbol{H}^H\boldsymbol{P}\boldsymbol{H})^{-1}$ inverts the channel correlation matrix to project each user's signal into the null space of others. Finally, the columns of $\boldsymbol{V}$ are normalized to the unit norm to ensure consistent power levels across transmission directions.

B. Received Signal

The received signal of the k^{th} user can be given as

$$\begin{aligned} \mathbf{y}_k &\overset{(a)}{=} \boldsymbol{H}_k^H\mathbf{x} + \mathbf{n}_k \\ &\overset{(b)}{=} \boldsymbol{H}_k^H\boldsymbol{E}\boldsymbol{V}\mathbf{s} + \boldsymbol{H}_k^H\mathbf{c} + \mathbf{n}_k \\ &\overset{(c)}{=} \boldsymbol{H}_k^H\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}\mathbf{s} + \boldsymbol{H}_k^H\boldsymbol{E}_{\text{OD}}\boldsymbol{V}\mathbf{s} + \boldsymbol{H}_k^H\mathbf{c} + \mathbf{n}_k \\ &\overset{(d)}{=} \boldsymbol{H}_k^H\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}_k\mathbf{s}_k + \boldsymbol{H}_k^H\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{V}_{-k}\mathbf{s}_{-k} + \boldsymbol{H}_k^H\mathbf{c} + \mathbf{n}_k, \end{aligned} \tag{8}$$

where in (a), channel matrix $\boldsymbol{H}_k$ for user k is an $M \times N_d$ matrix, representing the complex channel coefficients between the BS and the user. $\boldsymbol{H}_k$ takes into account both azimuth and elevation angles with the UPA configuration. These coefficients capture the path loss, shadowing, and multipath fading effects inherent in wireless communication. We model the channel as a Rayleigh fading channel, where each coefficient is independently drawn from a complex Gaussian distribution with zero mean and variance that corresponds to the path loss and shadowing effects between the respective antennas. $\mathbf{n}_k \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_N)$ is the additive Gaussian noise with zero means and σ^2 variance for each element; in (b) and (c), the quantities $\mathbf{x}$ and $\boldsymbol{E}$ are substituted; in (d), $\boldsymbol{V}_{-k} = [\boldsymbol{V}_1, \dots, \boldsymbol{V}_{k-1}, \boldsymbol{V}_{k+1}, \dots, \boldsymbol{V}_K]$ and $\mathbf{s}_{-k} = [\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_K]^H$ are used. At the k^{th} user, the resultant SINR, represented as user QoS, is given as

$$\xi_k(\boldsymbol{P}|\boldsymbol{H}) = \frac{P_0}{KN_d} \boldsymbol{H}_k^H \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{V}_k \boldsymbol{V}_k^H \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{H}_k \boldsymbol{\Sigma}_k^{-1}, \tag{9}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_k &= \mathrm{Cov}\Big(\boldsymbol{H}_k^H \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{V}_{-k} \mathbf{s}_{-k} + \boldsymbol{H}_k^H \mathbf{c} + \mathbf{n}_k\Big) & (10)\\ &= \frac{P_0}{KN_d} \boldsymbol{H}_k^H \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{V}_{-k} \boldsymbol{V}_{-k}^H \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{H}_k + \frac{P_0}{KN_d} \boldsymbol{H}_k^H \boldsymbol{G} \boldsymbol{H}_k \\ &\quad + \bar{\sigma}_c^2 \boldsymbol{H}_k^H \boldsymbol{H}_k + \sigma^2 \mathbf{I}. & (11) \end{aligned}$$

Thus, the resulting rate can be defined as $R_k(\boldsymbol{P}|\boldsymbol{H}) = \log_2 |\mathbf{I} + \xi_k(\boldsymbol{P}|\boldsymbol{H})|$.

C. Energy Efficiency for Fully Digital Precoding

In our model, we address the dynamics of wireless channels by considering them as time-varying across discrete time slots. For each time slot *t*, we assume the channel conditions remain constant. As [26], [27], we adopt a finite state Markov chain approach to represent the time-varying nature of the channel. This model allows the channel to transition among a predefined set of states, with each state corresponding to a distinct channel condition within a time slot. Through this approach, the ergodic nature of the channel is captured, where the state of the channel at any given time slot *t* is represented by one of the possible states in the Markov chain. Let $\mathcal{H} = \{\boldsymbol{H}^{(1)}, \dots, \boldsymbol{H}^{(|\mathcal{H}|)}\}$ denote the states in the Markov chain. The transition probability between channel states is fixed and unknown. Therefore, the resultant energy efficiency can be expressed as the ratio of the sum rate over the total power incurred in the transmission at time slot *t* as

$$\eta(t) = \frac{\sum_k R_k(t)}{C_{\text{tot}}(\boldsymbol{P}(t), \boldsymbol{H}(t))}. \tag{12}$$

If the CSI variations are Markov, the resultant SINR process will also be Markov.

IV. Hybrid Precoding Model

In the hybrid precoding model utilizing subarray-based hybrid beamforming, the BS is equipped with M antennas, which are uniformly distributed across KN_d subarrays. Each subarray is connected to a single RF chain, with each data stream associated with its own RF chain and corresponding subarray at the BS. Each user, equipped with N antennas, employs analog combining to receive N_d data streams. Consequently, the BS transmits a total of KN_d data streams via KN_d RF chains during each time slot [22].

A. Transmitted Signal

We assume that the dynamic communication process occurs within a total time duration of T, which is divided into multiple time intervals of duration ΔT. We employ a uniform planar array (UPA) for beamforming. In this setup, power allocation is explicitly represented for each antenna, user, and data stream in both the transmitted and received signals. Let $\mathbf{P'} \in \mathbb{C}^{M \times (K \times N_d)}$ denote the power allocation matrix for the hybrid precoding. The total transmit power constraint P_0' is enforced by ensuring that the sum of all elements of $\mathbf{P'}$ does not exceed the total power limit:

$$C'_{\text{tot}}(\mathbf{P'}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{N_d} p_{m,k,d} \leq P_0'. \tag{13}$$

$p_{m,k,d}$ is linked to each antenna, user, and data stream because shared RF chains in hybrid precoding require careful power distribution across streams and users within subarrays, compensating for limited control at the RF level. Fully digital precoding allocates power as p_m per antenna, as each antenna has its own RF chain, allowing independent data stream control at the digital level.

At the time $t \in [0, T]$, the signal transmitted from the BS can be given as [29],

$$\mathbf{x}(t) = \boldsymbol{F}_{\text{RF}} \sum_{k=1}^{K} \sum_{d=1}^{N_d} \mathbf{f}_{\text{BB},k,d} \sqrt{P_{k,d}} s_{k,d}(t) \Pi\left(\frac{t - 0.5 T_s}{T_s}\right), \tag{14}$$

where $\boldsymbol{F}_{\text{RF}} \in \mathbb{C}^{M \times (K \times N_d)}$ is the block diagonal RF precoder matrix, which applies RF precoding to each subarray associated with the N_d data streams of K users. The BB precoding vector $\mathbf{f}_{\text{BB},k,d} \in \mathbb{C}^{(K \times N_d) \times 1}$ is applied to the *d*-th data stream of the *k*-th user, and the power allocated to this stream is represented by $P_{k,d} = \sum_m^M p_{m,k,d}$, ensuring that each data stream is appropriately scaled. The data symbol $s_{k,d}(t)$ is the complex modulated signal for the *d*-th stream of the *k*-th user at time *t*, which is shaped by a rectangular pulse $\Pi(\frac{t-0.5T_s}{T_s})$ that defines the time window of the transmission. T_s is the duration of the symbol being transmitted. The transmitted signal $\mathbf{x}(t) \in \mathbb{C}^{M \times 1}$ combines these components, ensuring that each data stream is independently processed and transmitted through the corresponding antennas, with the overall transmission governed by the allocated power and the precoding strategy. The signal is a combination of BB and RF components, considering the power allocation across the precoding vectors.

B. Received Signal

For the *k*-th user ($k = 1, 2, \dots, K$), at time *t*, an RF combiner $\mathbf{w}_{\text{RF},k} \in \mathbb{C}^{N \times 1}$ is applied to receive the signal from the BS. The received signal can be expressed as:

$$\begin{aligned} y_k(t) = \mathbf{w}_{\text{RF},k}^H \Bigg[\boldsymbol{H'}_k^H \Bigg(\mathbf{x}_k(t) + \sum_{k' \in \mathcal{K}, k' \neq k} \mathbf{x}_{k'}(t) + \boldsymbol{F}_{\text{RF}} \mathbf{c}_{\text{R}}(t) \Bigg) \\ + \mathbf{z}_{\text{r},k}(t) \Bigg] + e_{\text{V},k}(t), \end{aligned} \tag{15}$$

where, $\mathbf{x}_k(t) = \boldsymbol{F}_{\text{RF}} \sum_{d=1}^{N_d} \mathbf{f}_{\text{BB},k,d} \sqrt{P_{k,d}} s_{k,d}(t) \Pi(\frac{t-0.5T_s}{T_s}) \in \mathbb{C}^{M \times 1}$, and similarly for $\mathbf{x}_{k'}(t)$, which represents the signal from other users k' causing interference. Hybrid precoding should consider inter-user interference because it relies on a combination of analog and digital precoding, where the analog component does not completely eliminate interference due to its limited degrees of freedom and the shared use of RF chains among multiple antennas. So the interference from other users must be explicitly accounted for to ensure accurate modeling. In contrast, fully digital precoding typically employs ZF techniques, which can effectively nullify inter-user interference. This makes the explicit consideration of such interference unnecessary in fully digital precoding models. $\boldsymbol{H'}_k = \beta_k \sqrt{NM} \mathbf{a}_N(\theta_k, \phi_k) \mathbf{a}_M^H(\theta_k, \phi_k)$ is the channel matrix for hybrid precoding, where $\mathbf{a}_N(\theta_k, \phi_k)$ is the array response vector of the receiver antenna array, and $\mathbf{a}_M^H(\theta_k, \phi_k)$ is the array response vector of the transmitter antenna array. Here, θ_k and ϕ_k represent the azimuth and elevation angles of the k-th user, respectively. The fading coefficient $\beta_k = \bar{d}_k^{-1} e^{j \frac{2\pi}{w_l} \bar{d}_k}$ accounts for the propagation effects, with $\bar{d}_k = \sqrt{d_k^2 + h^2}$ being the propagation distance between the k-th user and the BS, where h denotes the BS height, and d_k is the horizontal distance between the BS and the user. w_l represents the wavelength of the transmitted signal. The term $\mathbf{z}_{\mathrm{r},k}(t) \in \mathbb{C}^{N \times 1}$ denotes the additive white Gaussian noise (AWGN) vector at the receiver, while $\mathbf{c}_{\mathrm{R}}(t)$ is a zero-mean Gaussian random variable representing the transmitter HWI, including oscillator noise, amplifier noise, and nonlinearities in the DACs and ADCs. The final term, $e_{\mathrm{V},k}(t)$, represents the receiver HWI, modeled as a zero-mean Gaussian random process with variance σ_c^2. The modeling of HWI as a Gaussian random variable in hybrid precoding is practical because the aggregate impairments follow a Gaussian distribution due to the central limit theorem, especially since hybrid precoding involves fewer RF chains shared among multiple antennas. The BB combiner is neglected in the received signal because the RF combiner handles most of the beamforming and interference management, efficiently utilizing limited RF chains and reducing processing complexity.

C. RF Beamformers and MMSE BB Precoders

We assume that users can accurately acquire the direction of the BS, which is a practical and reasonable assumption given current technology such as angle-of-arrival (AoA) estimation [30] and MUSIC (Multiple Signal Classification) [31]. This allows us to simplify the model. The k-th block of the RF precoder matrix for the BS, associated with the d-th data stream, is defined as:

$$\mathbf{f}_{\mathrm{RF},k,d} = \sqrt{\frac{M}{N_d}} [\mathbf{a}_M(\theta_k, \phi_k)]_{\frac{(k-1)M}{K} + \frac{(d-1)M}{KN_d} + 1 : \frac{(k-1)M}{K} + \frac{dM}{KN_d}, :}, \tag{16}$$

where (θ_k, ϕ_k) are the azimuth and elevation angles of the k-th user. This subscript notation defines the range of antenna indices allocated to the d-th data stream of the k-th user, starting from $\frac{(k-1)M}{K} + \frac{(d-1)M}{KN_d} + 1$ and ending at $\frac{(k-1)M}{K} + \frac{dM}{KN_d}$. This configuration allows each data stream d to have its own RF beamforming pattern within the k-th user's allocated portion of the BS antenna array. The RF combiner for the k-th user is defined as:

$$\mathbf{w}_{\mathrm{RF},k} = \sqrt{N} \mathbf{a}_N(\theta_k, \phi_k), \tag{17}$$

which can imply that $\mathbf{w}_{\mathrm{RF},k}^H \boldsymbol{H'}_k = \sqrt{MN} \beta_k \mathbf{a}_M^H(\theta_k, \phi_k)$.

The MMSE BB precoder is designed to mitigate multi-user interference and is defined as:

$$\boldsymbol{F}_{\mathrm{BB}} = f_{\mathrm{R}} \widetilde{\boldsymbol{F}}_{\mathrm{BB}}, \tag{18}$$

where $\boldsymbol{F}_{\mathrm{BB}} = [\mathbf{f}_{\mathrm{BB},1,1}, \mathbf{f}_{\mathrm{BB},1,2}, \ldots, \mathbf{f}_{\mathrm{BB},K,N_d}] \in \mathbb{C}^{(K \times N_d) \times (K \times N_d)}$ represents the MMSE BB precoder matrix for all $K \times N_d$ data streams. $f_{\mathrm{R}} = \sqrt{\frac{1}{\|\boldsymbol{F}_{\mathrm{RF}} \widetilde{\boldsymbol{F}}_{\mathrm{BB}}\|_F^2}}$ is a normalization factor to normalize the precoder to unit power, ensuring a consistent basis for applying power constraints at the signal level.

The effective BB precoder $\widetilde{\boldsymbol{F}}_{\mathrm{BB}} = (\boldsymbol{H'}_{\mathrm{R}}^H \boldsymbol{H'}_{\mathrm{R}} + \varpi \operatorname{diag}(\boldsymbol{H'}_{\mathrm{R}}^H \boldsymbol{H'}_{\mathrm{R}}) + \sigma^2 \boldsymbol{F}_{\mathrm{RF}}^H \boldsymbol{F}_{\mathrm{RF}})^{-1} \boldsymbol{H'}_{\mathrm{R}}^H$, where $\boldsymbol{H'}_{\mathrm{R}} = \boldsymbol{A}_{\mathrm{R}}^H \boldsymbol{F}_{\mathrm{RF}}$ and $\boldsymbol{A}_{\mathrm{R}} = [\mathbf{a}_M(\theta_1, \phi_1), \mathbf{a}_M(\theta_2, \phi_2), \ldots, \mathbf{a}_M(\theta_K, \phi_K)]$, accounts for the effective channel after RF beamforming. The parameters $\varpi, \varsigma \ll 1$, are the HWI factors, modeling the impact of practical hardware limitations on the system performance.

D. Energy Efficiency for Hybrid Precoding

At time t, the received signal for the k-th user can be expressed as:

$$\begin{aligned} y_k(t) = \mathbf{w}_{\mathrm{RF},k}^H \Bigg[& \boldsymbol{H'}_k^H \boldsymbol{F}_{\mathrm{RF}} \sum_{d=1}^{N_d} \mathbf{f}_{\mathrm{BB},k,d} \sqrt{P_{k,d}} s_{k,d}(t) \\ & + \sum_{k' \in \mathcal{K}, k' \neq k} \boldsymbol{H'}_k^H \boldsymbol{F}_{\mathrm{RF}} \sum_{d=1}^{N_d} \mathbf{f}_{\mathrm{BB},k',d} \sqrt{P_{k',d}} \mathbf{s}_{k',d}(t) \\ & + \boldsymbol{H'}_k^H \boldsymbol{F}_{\mathrm{RF}} \mathbf{c}_{\mathrm{R}}(t) + \mathbf{z}_{\mathrm{r},k}(t) \Bigg] + e_{\mathrm{V},k}(t), \end{aligned} \tag{19}$$

where $\mathbf{z}_{\mathrm{r},k}(t)$ has zero mean and variance σ^2, and $e_{\mathrm{V},k}(t)$ has zero mean and variance $\varsigma \mathbb{E}\{\tilde{y}_k(t) \tilde{y}_k^H(t)\}$ with $\tilde{y}_k(t) = y_k(t) - e_{\mathrm{V},k}(t)$. The covariance of the useful signal and interference is given by:

$$\mathbb{E}\Big\{\tilde{y}_k(t) \tilde{y}_k^H(t)\Big\} = \sum_{d=1}^{N_d} \hat{P}_{k,d} + \sum_{k' \in \mathcal{K}, k' \neq k} \sum_{d=1}^{N_d} \hat{P}_{k,k',d} + M\sigma^2, \tag{20}$$

where the term $\hat{P}_{k,d} = P_{k,d} \|\mathbf{w}_{\mathrm{RF},k}^H \boldsymbol{H'}_k^H \boldsymbol{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB},k,d}\|^2$ is the desired signal power for the d-th data stream of the k-th user. $\hat{P}_{k,k',d} = P_{k',d} \|\mathbf{w}_{\mathrm{RF},k}^H \boldsymbol{H'}_k^H \boldsymbol{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB},k',d}\|^2$ is the interference power from the d-th data stream of the k'-th user at the k-th user. The noise power is captured by $M\sigma^2$. The SINR for the k-th user at time t is then given by:

$$\xi_k'(t) = \frac{\sum_{d=1}^{N_d} \hat{P}_{k,d}}{\sum_{k' \in \mathcal{K}, k' \neq k} \sum_{d=1}^{N_d} \hat{P}_{k,k',d} + M\sigma^2 + \varsigma \mathbb{E}\{\tilde{y}_k(t) \tilde{y}_k^H(t)\}}, \tag{21}$$

where the numerator represents the power of the desired signal, and the denominator includes the interference from other users' data streams, noise, and hardware impairment noise. The data rate for the *k*-th user at time *t* can be expressed as $R'_k(t) = \log_2(1 + \xi'_k(t))$. Therefore, the energy efficiency for the hybrid precoding at time *t* is given by:

$$\eta'(t) = \frac{\sum_{k=1}^{K} R'_k(t)}{C'_{\text{tot}}\big(\boldsymbol{P'}(t)\big)}, \tag{22}$$

where $C'_{\text{tot}}(\boldsymbol{P'}(t))$ represents the total power consumption at time *t*, considering all users and their respective data streams.

V. Problem Formulation

In this work, our goal is to maximize the long-term energy efficiency of transmissions via efficient power allocations. The system operates in discrete time slots, and the MDP framework addresses the NP-hard and non-convex nature of the problem of efficient power allocation in each time slot. This framework assumes that channel information changes according to a finite state Markov chain with unknown transition probabilities. The MDP problem aims to find an efficient strategy for power allocation that maximizes long-term energy efficiency while adhering to the dynamic QoS constraint $\bar{\xi}_k(t)$, a total power constraint, and per-antenna power constraints, thereby enhancing decision-making in the face of uncertain and varying channel conditions. Therefore, for the fully digital precoding, we formulate the problem as:

$$\max_{\boldsymbol{P}(t)} \sum_{\tau=t}^{\infty} \gamma^{\tau-t}\eta(t) \tag{23}$$

$$\begin{aligned} &\text{subject to} \\ &C_{\text{tot}}(\boldsymbol{P}(t), \boldsymbol{H}(t)) \leq P_0, \\ &p_m(t) \leq \bar{p}_{\max}, \forall m \in \mathcal{M}, \\ &\xi_k(\boldsymbol{P}(t)|\boldsymbol{H}(t)) \geq \bar{\xi}_k(t), \forall k \in \mathcal{K}, \end{aligned} \tag{24}$$

where $\bar{p}_{\max}$ denotes the per-antenna power upper constraint for the digital precoding. γ is the discount factor.

For the hybrid precoding, we just need to substitute the corresponding variables as:

$$\max_{\boldsymbol{P'}(t)} \sum_{\tau=t}^{\infty} \gamma^{\tau-t}\eta'(t) \tag{25}$$

$$\begin{aligned} &\text{subject to} \\ &C'_{\text{tot}}\big(\boldsymbol{P'}(t)\big) \leq P'_0, \\ &p_{m,k,d}(t) \leq \bar{p}'_{\max}, \forall m \in \mathcal{M}, k \in \mathcal{K}, d \in \{1, \ldots, N_d\}, \\ &\xi'_k(t) \geq \bar{\xi}_k(t), \forall k \in \mathcal{K}, \end{aligned} \tag{26}$$

where $\bar{p}'_{\max} = \frac{P'_0}{MKN_d}$ denotes the per-antenna-user-stream power constraint for hybrid precoding.

VI. Solution via Reinforcement Learning

A. Motivation of REDQ

The optimization problems outlined for fully digital and hybrid precoding and efficient decision-making policies are derived in dynamic environments where the channel characteristics vary with time and are not known in advance. Considering the time-varying nature of channels across different time slots, our approach leverages CSI correlations through a Markov process. Consequently, the power allocation challenges in fully digital and hybrid precoding scenarios are aptly formulated as MDP problems. Reinforcement learning emerges as an apt methodology to enhance long-term energy efficiency, owing to its intrinsic strength in managing sequential decision-making tasks within these dynamic environments.

Furthermore, in the analog power allocation problem, the action spaces are real-valued, so it is not appropriate for the training algorithm to use a reinforcement learning algorithm that applies to discrete actions, such as the deep *Q* network (DQN) algorithm. Therefore, we use a REDQ algorithm [32] to train the process. REDQ is a cutting-edge reinforcement learning technique that improves sample efficiency in continuous control tasks. It achieves this by using a combination of multiple *Q*-function approximators, a high Update-To-Data (UTD) ratio, and a strategy that minimizes overestimation bias through targeted *Q*-value updates. The UTD ratio refers to the number of learning updates performed for each new data sample collected from the environment. A higher UTD ratio means the algorithm makes more updates per data sample, potentially leading to faster learning and better utilization of each experience. REDQ can achieve a high UTD ratio without a model, prevents overfitting, and maintains stability. This is realized through an ensemble of *Q*-functions and randomized selection for target calculation, which diversifies estimates and reduces over-reliance on single *Q*-function biases. Soft updates for target parameters ensure gradual, stable learning signals, while entropy regularization promotes exploration, preventing premature convergence to suboptimal policies. This combination of strategies allows REDQ to efficiently leverage each data sample for multiple updates, enhancing learning efficiency and robustness in complex environments without requiring synthetic data generation.

B. Reinforcement Learning Settings for Fully Digital Precoding

1) States, Actions, and Rewards for Fully Digital Precoding: In the above system dynamics, let $\mathbf{s}(t)$ represent the state at time *t*, defined as the current CSI for that slot, i.e., $\mathbf{s}(t) = \boldsymbol{H}(t) \in \mathcal{H}$. The system's action corresponds to analog power control, expressed as $\mathbf{a}(t) = \boldsymbol{P}(t)$. For a BS with *M* antennas, the range of possible actions is infinite; however, not all actions are valid. Valid actions are those that meet the power constraint in (7). Specifically, the total power $C_{\text{tot}}(\boldsymbol{P}(t), \boldsymbol{H'}(t))$ depends on the normalized precoder $\boldsymbol{V}(t)$, ensuring that only feasible power allocations are chosen. To reduce the valid action set, we approximate the total power constraint as

$$\begin{aligned} &\frac{P_0}{KN_d}\operatorname{tr}\Big(\boldsymbol{P}^{\frac{1}{2}}(t)\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{P}^{\frac{1}{2}}(t)\Big) + \frac{P_0}{KN_d}\operatorname{tr}(\boldsymbol{G}(t)) + \bar{\sigma}_c^2\operatorname{tr}(\mathbf{I}) \approx \\ &\quad \frac{P_0}{KN_d}\operatorname{tr}\Big(\boldsymbol{P}^{\frac{1}{2}}(t)\mathbb{E}\big\{\boldsymbol{V}_R\boldsymbol{V}_R^H\big\}\boldsymbol{P}^{\frac{1}{2}}(t)\Big) + \frac{P_0}{KN_d}\operatorname{tr}(\boldsymbol{G}(t)) + \bar{\sigma}_c^2\operatorname{tr}(\mathbf{I}) \\ &\quad = \frac{P_0}{M}\operatorname{tr}(\boldsymbol{P}(t)) + \frac{P_0}{KN_d}\operatorname{tr}(\boldsymbol{G}(t)) + \bar{\sigma}_c^2\operatorname{tr}(\mathbf{I}) \leq P_0, \end{aligned} \tag{27}$$

where $\boldsymbol{V}_R$ is any random orthonormal precoder. The equality on the right follows from [33, Lemma 1]. Further, to ensure all data streams are served, at least KN_d power allocations should be non-zero (i.e., $p_{m_d} > 0$ for each required data stream d), thereby excluding any actions with fewer than KN_d non-zero allocations from the action space. To further reduce the number of possibilities in the power allocation process, we approximate the minimum transmission QoS constraint at a given time slot t for the k-th user as

$$\begin{aligned}
&\xi_k(\boldsymbol{P}(t)|\boldsymbol{H}(t))\\
&\overset{(a)}{\approx} \frac{\frac{P_0}{KN_d}\operatorname{tr}\Big(\boldsymbol{P}(t)\boldsymbol{V}_k\boldsymbol{V}_k^H\Big)}{\frac{P_0}{KN_d}\operatorname{tr}\Big(\boldsymbol{P}(t)\boldsymbol{V}_{-k}\boldsymbol{V}_{-k}^H\Big)+\frac{P_0}{KN_d}\operatorname{tr}(\boldsymbol{G}(t))+\bar{\sigma}_c^2+\sigma^2}\\
&\overset{(b)}{\approx} \frac{\frac{P_0}{KM}\operatorname{tr}(\boldsymbol{P}(t))}{\frac{P_0(KN_d-1)}{KM}\operatorname{tr}(\boldsymbol{P}(t))+\frac{P_0}{KN_d}\operatorname{tr}(\boldsymbol{G}(t))+\bar{\sigma}_c^2+\sigma^2}\\
&= \frac{1}{(KN_d-1)\ +M\frac{\operatorname{tr}(\boldsymbol{G}(t))+\frac{KN_d}{P_0}(\bar{\sigma}_c^2+\sigma^2)}{\operatorname{tr}(\boldsymbol{P}(t))}},
\end{aligned} \tag{28}$$

Step (a) uses the channel hardening effect, which indicates that in massive MIMO systems, as the number of antennas M increases, the channel vector $\boldsymbol{H}_k$ becomes nearly deterministic, which can be represented as $\boldsymbol{H}_k\boldsymbol{H}_k^H \approx \mathbf{I}_M$. Thus, $\operatorname{tr}(\boldsymbol{H}_k^H\boldsymbol{P}^{1/2}(t)\boldsymbol{V}_{-k}\boldsymbol{V}_{-k}^H\boldsymbol{P}^{1/2}(t)\boldsymbol{H}_k)$ can be approximated by $\operatorname{tr}(\boldsymbol{P}(t)\boldsymbol{V}_{-k}\boldsymbol{V}_{-k}^H)$. Step (b) further simplifies the expression by assuming $\boldsymbol{V}_k$ is a random orthonormal matrix. For such matrices, the expectation $\mathbb{E}\{\boldsymbol{V}_k\boldsymbol{V}_k^H\}$ is $\mathbf{I}_M$. Therefore,

$$\operatorname{tr}\Big(\boldsymbol{P}(t)\boldsymbol{V}_k\boldsymbol{V}_k^H\Big) \approx \frac{\operatorname{tr}(\boldsymbol{P}(t))}{M}, \tag{29}$$

and similarly,

$$\operatorname{tr}\Big(\boldsymbol{P}(t)\boldsymbol{V}_{-k}\boldsymbol{V}_{-k}^H\Big) \approx \frac{\operatorname{tr}(\boldsymbol{P}(t))(KN_d-1)}{M}. \tag{30}$$

This approximation comes from the property of random orthonormal matrices where each entry's contribution is uniformly spread out across the dimensions. To ensure the SINR meets the minimum QoS requirement $\bar{\xi}_k(t)$, we derive a lower bound on the transmission power. For ZF precoding, this condition can be written as

$$\frac{1}{(KN_d-1)\ +M\frac{\operatorname{tr}(\boldsymbol{G}(t))+\frac{KN_d}{P_0}(\bar{\sigma}_c^2+\sigma^2)}{\operatorname{tr}(\boldsymbol{P}(t))}} \geq \bar{\xi}_k(t), \tag{31}$$

which simplifies to

$$\operatorname{tr}(\boldsymbol{P}(t)) \geq \frac{M\bar{\xi}_k(t)\Big[\operatorname{tr}(\boldsymbol{G}(t))+\frac{KN_d}{P_0}\big(\bar{\sigma}_c^2+\sigma^2\big)\Big]}{1-\bar{\xi}_k(t)(KN_d-1)}. \tag{32}$$

Therefore, the minimum transmission power required to meet the QoS constraint is given by

$$\bar{P}_{\min}(t) = \frac{M\bar{\xi}_k(t)\Big[\operatorname{tr}(\boldsymbol{G}(t))+\frac{KN_d}{P_0}\big(\bar{\sigma}_c^2+\sigma^2\big)\Big]}{1-\bar{\xi}_k(t)(KN_d-1)}. \tag{33}$$

Thus, the constrained action space can be,

$$\bar{\boldsymbol{P}}_M = \left\{\begin{pmatrix} p_1\\ \vdots\\ p_M\end{pmatrix} : \begin{array}{c}\bar{P}_{\min}(t) \leq tr(\boldsymbol{P}(t)) \leq P_0,\\ 0 < p_{m_d} < \bar{p}_{\max}, \forall d = 1,\ldots,KN_d.\end{array}\right\}. \tag{34}$$

The action space reduction improves the quality and efficiency of the learning process by guiding the agent toward feasible actions, leading to faster convergence, more consistent rewards, and reliable compliance with constraints. However, this reduction focuses on eliminating infeasible actions rather than reducing the dimensionality of the action space or the complexity of evaluating each action. As such, while action space reduction aids learning efficiency, it does not directly lower computational complexity.

The reward evaluating the action is defined as energy efficiency divided by the QoS, i.e.,

$$\begin{aligned}
r(\mathbf{s}(t),\mathbf{a}(t)) = &\ \frac{\eta(t)}{\sum_{k\in\mathcal{K}}\big|\xi_k(\boldsymbol{P}(t)|\boldsymbol{H}(t))-\bar{\xi}_k(t)\big|},\\
&-\lambda_1[\max(0,\operatorname{tr}(\boldsymbol{P}(t))-P_0)]^2\\
&-\lambda_2\left[\max\left(0,\sum_m(p_m(t)-\bar{p}_{\max})\right)\right]^2\\
&-\lambda_3\big[\max\big(0,\bar{P}_{\min}(t)-\operatorname{tr}(\boldsymbol{P}(t))\big)\big]^2,
\end{aligned} \tag{35}$$

where $\lambda_n > 0, n \in \{1,2,3\}$ denote the penalty parameter to determine the penalty magnitude. $|\cdot|$ ensures that the resulting SINR does not achieve values far from $\bar{\xi}_k(t)$. Here, the learner seeks the optimum action $\mathbf{a}(t)$ based on the previous observation $\boldsymbol{H}(t-1) = s(t-1)$ by interactively making sequential decisions and observing the corresponding costs. In this way, the agent learns the best action policy against the random Markov chain transitions. Let the policy function be $\pi : \mathcal{H} \to \boldsymbol{P}$, which maps a state to an action. Under policy $\pi(\cdot)$, the power allocation is carried out via action $a(t+1) = \pi(\mathbf{s}(t))$, dictating the allocation policy at time $t+1$. For the reward $r_\pi(\mathbf{s}(t)) = r(\mathbf{s}(t),\pi(\mathbf{s}(t)))$, power consumption performance is measured through the state value function as $V_\pi(\mathbf{s}(t)) = \sum_{\tau=t}^{\infty}\gamma^{\tau-t}r_\pi(\mathbf{s}(t))$, which is the total average cost incurred over an infinite time horizon. The objective of this paper is to find the efficient policy π^* such that the average cost of any state is maximized $\pi^* = \arg\max_\pi V_\pi(\mathbf{S})$.

2) Analog and Discrete Power Allocation for Fully Digital Precoding: In analog power allocation for fully digital precoding, each antenna's power level is chosen from a continuous range $[0,\bar{p}_{\max}]$ and must satisfy both per-antenna and total power constraints $\bar{P}_{\min}(t)$ and P_0, as specified in (34).

For discrete power allocation, the power allocation set $\mathcal{P} = \{p^{(1)},\ldots,p^{(|\mathcal{P}|)}\}$ contains all possible power levels for per-antenna power allocation. Each power level p_m belongs to $\mathcal{P}$ for every antenna m, meaning that $p_m \in \mathcal{P}, \forall m$. The power allocation matrix $\boldsymbol{P} \in \mathcal{P}^M$ satisfies $0 \leq p^{(1)} \leq \cdots \leq p^{(|\mathcal{P}|)} = \bar{p}_{\max}$. The action space $\mathbf{a}_t = [a_{1t},\ldots,a_{Mt}]^T \in \mathcal{A}^M$ indexes the power levels chosen for each antenna, resulting in an action space of $|\mathcal{P}|^M$, which grows exponentially with the number of antennas. To manage this, an action coding scheme is proposed [34]. In this scheme, the action index a_{mt} is first encoded into a real number between 0 and 1 as: $a_m(t) = \frac{a_{mt}}{|\mathcal{P}|}$, which is then mapped to the new action space $\mathcal{A}_{\text{new}} = [0,1]$. This encoding scheme reduces the matrix-valued power allocation action to a single scalar, $\mathbf{a}(t) = [a_1(t),\ldots,a_M(t)]^T \in \mathcal{A}_{\text{new}}^M$.

To decode the real number into a valid power allocation action, the real-valued power allocation $a_m(t)$ for antenna m must be converted back to the discrete range of the original power set $\mathcal{P}$. During this decoding process, the binary representation $\mathbf{b}_{|\mathcal{P}|}(n) \in \{0,1\}^{|\mathcal{P}|}$, a $|\mathcal{P}| \times 1$ vector, is used for each power allocation index $n = a_m(t) \times |\mathcal{P}|$, where $a_m(t)$ is the encoded power allocation for antenna m.

Once the learning agent proposes a power allocation, we utilize a supplementary constraint-adjustment mechanism to strictly enforce the total and per-antenna power constraints. Although the reward includes penalty terms to discourage violations, this mechanism is necessary to ensure immediate compliance, especially during early exploration when actions may exceed limits. Using the binary representation, the valid power allocation action is obtained by flipping the least significant bits (LSBs) to satisfy the power constraint $\mathbf{1}^T_{|\mathcal{P}|}\mathbf{b}_{|\mathcal{P}|}(a_m(t) \times |\mathcal{P}|) = B_m(t) \leq \bar{p}_{\max}$. Each 1 in the m^{th} position of $\mathbf{b}_{|\mathcal{P}|}(a_m(t) \times |\mathcal{P}|)$ signifies the power allocation action of the m^{th} antenna. If $|B_m(t) - \bar{p}_{\max}| > 0$, the process involves flipping the 1's entries to 0 from the LSBs until the power level does not exceed $\bar{p}_{\max}$, with the LSBs of $\mathbf{b}_{|\mathcal{P}|}(n)$ representing the smallest power level. Once the per-antenna power constraint is satisfied, the sum of power levels across all antennas must not exceed the total power budget P_0. If the total power level exceeds P_0, adjustments are made by modifying the binary representation for KN_d data streams assigned to all users from the LSBs until the total power level fits within the limit P_0. Similarly, adjustments ensure the minimum transmission power constraint $\bar{P}_{\min}(t)$ is met.

C. Reinforcement Learning Settings For Hybrid Precoding

1) States, Actions, and Rewards for Hybrid Precoding: Similarly, the state $\mathbf{s}(t) = \boldsymbol{H}'(t)$ is defined as the set of the current CSI for all users. The action $\mathbf{a}(t) = \boldsymbol{P}'(t)$ involves allocating power to different antennas and data streams across all users. The reward is the energy efficiency divided by the QoS as

$$
\begin{aligned}
r(\mathbf{s}(t), \mathbf{a}(t)) = &\frac{\eta'(t)}{\sum_{k,d}\left|\xi'_{k,d}(t) - \bar{\xi}_{k,d}(t)\right|} \\
&- \lambda'_1\left[\max\left(0, C'_{\text{tot}}(\boldsymbol{P}'(t)) - P'_0\right)\right]^2 \\
&- \lambda'_2\left[\max\left(0, \sum_{m,k,d}\left(p_{m,k,d}(t) - \bar{p}'_{\max}\right)\right)\right]^2 \\
&- \lambda'_3\left[\max\left(0, \sum_{k,d}\left(\bar{\xi}_{k,d}(t) - \xi'_{k,d}(t)\right)\right)\right]^2,
\end{aligned}
\tag{36}
$$

where $\lambda'_n > 0, n \in \{1,2,3\}$ denote the penalty parameter for hybrid precoding.

2) Analog and Discrete Power Allocation for Hybrid Precoding: For analog power allocation in hybrid precoding, the power allocated to each antenna, grouped under an RF chain associated with a user's data stream, is selected from a continuous, real-valued action space defined as $[0, \bar{p}'_{\max}]$, where $\bar{p}'_{\max}$ represents the maximum allowable power for each antenna within a particular RF chain, user, and data stream.

For discrete power allocation, the set $\mathcal{P}' = \{p^{(1)}, \ldots, p^{(|\mathcal{P}'|)}\}$ defines possible power levels for each antenna, where $0 \leq p^{(1)} \leq \cdots \leq p^{(|\mathcal{P}'|)} = \bar{p}'_{\max}$. In hybrid precoding, each RF chain supports a subarray of antennas rather than each antenna having its own dedicated RF chain, allowing multiple antennas to be controlled together by one RF chain. However, power allocation decisions are still made at the level of each individual antenna to enable flexible and precise control of the transmit power. Consequently, the complete power allocation tensor $\boldsymbol{P}'$ is an element of the set $\mathcal{P}'^M$. The action space simplifies to M because, although the number of RF chains is reduced to match the number of data streams, which is $K \times N_d$, each of the M antennas still requires an independent power allocation. In hybrid precoding, each RF chain controls a subarray of $\frac{M}{K \times N_d}$ antennas. Thus, while the system uses only $K \times N_d$ RF chains, power allocation must still be configured for each antenna within these subarrays to achieve the required flexibility in transmit power control. For the action space, the action vector $\mathbf{a}_t$ is thus defined as an element of $\mathcal{A}^M$, where each $a_{m,t}$ corresponds to an index in $\mathcal{P}'$, representing the power allocation for each individual antenna rather than solely at the RF chain level.

The action coding scheme is also used to manage the complexity by encoding each power level index $a_{mkd,t}$ as a real number $a_{mkd}(t) = \frac{a_{mkd,t}}{|\mathcal{P}'|} \in [0,1]$. Decoding maps these real values back to $\mathcal{P}'$ using a binary representation $\mathbf{b}_{|\mathcal{P}'|}(n)$ for each index $n = a_{mkd}(t) \times |\mathcal{P}'|$. For the supplementary constraint-adjustment mechanism, we adjust the power allocation action to satisfy the constraint $\mathbf{1}^T_{|\mathcal{P}'|}\mathbf{b}_{|\mathcal{P}'|}(a_{mkd}(t) \times |\mathcal{P}'|) = B_{mkd}(t) \leq \bar{p}'_{\max}$. If a power violation occurs, meaning $|B_{mkd}(t) - \bar{p}'_{\max}| > 0$, we mitigate this by systematically flipping the entries from 1 to 0, starting with the LSBs, until the power level falls within the allowable limits defined by $\bar{p}'_{\max}$ and the total power constraint P'_0.

D. Training Process of REDQ

The training process of the REDQ algorithm enhances reinforcement learning (RL) models' sample efficiency and performance, especially in continuous action spaces. REDQ uses multiple *Q*-functions to estimate the value of actions, which helps in reducing overestimation bias. A high UTD ratio ensures frequent updates of *Q*-functions relative to the number of new samples collected, improving learning stability and speed. The in-target minimization technique further refines the policy by selecting the minimum value among *Q*-functions during updates, enhancing robustness against value overestimation. This combination of techniques makes REDQ particularly effective for complex, dynamic environments. Here's a detailed explanation of the REDQ training process:

1) Initialization: The REDQ algorithm starts by initializing the policy network with parameters Θ, which dictates the strategy for selecting actions based on the current state. Concurrently, it initializes parameters for J *Q*-functions, Φ_j for $j = 1, \ldots, J$, which estimates the expected returns of taking certain actions from given states. The target

Q-function parameters, $\Phi_{\mathrm{targ},j}$, are also initialized to mirror the *Q*-function parameters, aiding in stabilizing the learning updates. A replay buffer $\mathcal{O}$ is established to store experience tuples $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$, where **s** is the state, **a** is the action taken, *r* is the received reward, and **s**′ is the subsequent state. This buffer serves as a reservoir of experiences for training the *Q*-functions and updating the policy.

2) Data Collection: In the data collection phase, the algorithm interacts with the environment to accumulate experiences. For each timestep *t*, the policy π_Θ selects an action $\mathbf{a}(t)$ based on the current state $\mathbf{s}(t)$, and the environment responds with a reward $r(t)$ and a new state $\mathbf{s}(t+1)$. These experiences are stored in the replay buffer $\mathcal{O}$, forming a dataset for training. This process is crucial for learning effective policies and *Q*-function estimates, as it provides the raw data from which the algorithm learns.

3) Updating Q-Functions: Updating the *Q*-functions is a critical step where the algorithm refines its estimates of expected returns. For each update iteration, a mini-batch $\mathcal{B}$ of experiences is sampled from $\mathcal{O}$, and a subset of *Q*-functions of size J' is randomly selected. For each experience in $\mathcal{B}$, the target *Q*-value *q* is computed as:

$$q = r + \gamma \min_{j \in J'} Q_{\Phi_{\mathrm{targ},j}}\left(\mathbf{s}', \tilde{\mathbf{a}}'\right) - \alpha \log \pi_\Theta\left(\tilde{\mathbf{a}}' | \mathbf{s}'\right), \quad (37)$$

where γ is the discount factor, $\tilde{\mathbf{a}}'$ is the action proposed by the policy for the next state **s**′, and α is the temperature parameter controlling the importance of the entropy term $\log \pi_\Theta(\tilde{\mathbf{a}}'|\mathbf{s}')$ for exploration. The *Q*-function parameters Φ_j are updated by minimizing the mean squared error between the *Q*-function's current estimates and the targets *q*, enhancing the accuracy of the *Q*-value predictions.

4) Policy Improvement: After updating the *Q*-functions, the algorithm refines the policy by optimizing the parameters Θ, aiming to maximize the expected rewards. This optimization is executed through gradient ascent on a well-defined objective function. This function is designed to prefer actions that are anticipated by the *Q*-functions to produce higher rewards, while also incorporating an entropy term to promote sufficient exploration. The entropy term is instrumental in preventing the policy from prematurely converging to suboptimal deterministic behaviors by encouraging exploration. For each *Q*-function indexed by *j*, the algorithm first adjusts its parameters Φ_j to better approximate the expected rewards. This adjustment is achieved by minimizing the mean squared error between the predicted *Q*-values and the computed target values *q* for all transitions in the mini-batch $\mathcal{B}$:

$$\text{Minimize: } \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{s},\mathbf{a},r,\mathbf{s}') \in \mathcal{B}} \left(Q_{\Phi_j}(\mathbf{s}, \mathbf{a}) - q\right)^2. \quad (38)$$

Subsequently, the target parameters $\Phi_{\mathrm{targ},j}$ for each *Q*-function are updated using a soft update rule:

$$\Phi_{\mathrm{targ},j} \leftarrow \rho \Phi_{\mathrm{targ},j} + (1-\rho)\Phi_j, \quad (39)$$

where ρ is a hyperparameter that controls the rate of updating the target parameters, typically set close to 1 to ensure gradual updates. Once the *Q*-function parameters and their corresponding target parameters have been updated, the algorithm proceeds to refine the policy parameters Θ. This refinement is conducted by applying gradient ascent to maximize the expected return, adjusted for exploration:

$$\Delta\Theta = \omega \nabla_\Theta \mathbb{E}_{\mathbf{s}\sim\mathcal{B}, \tilde{\mathbf{a}}\sim\pi_\Theta}\left[\frac{1}{J}\sum_{j=1}^{J} Q_{\Phi_j}(\mathbf{s}, \tilde{\mathbf{a}}) - \alpha \log \pi_\Theta(\tilde{\mathbf{a}}|\mathbf{s})\right], \quad (40)$$

where ω denotes the learning rate. This equation indicates that the update to the policy parameters Θ is directly proportional to the gradient of the expected return, as estimated by the ensemble of *Q*-functions, minus a term proportional to the policy's entropy to ensure exploration. This structured approach ensures that the policy is iteratively refined using the most up-to-date and stable estimates from the *Q*-functions, guiding the policy towards selecting actions that are expected to yield higher returns and thereby enhancing the agent's performance in the environment.

5) Iteration and Convergence: The algorithm iterates through the cycle of data collection, *Q*-function updating, and policy improvement, gradually enhancing the policy's performance. Let the maximum number of training steps be N_e. A key aspect of REDQ is its high UTD ratio, denoted by *D*, which dictates the number of updates performed for each batch of collected data. This high UTD ratio is instrumental in extracting significant learning value from each interaction, thereby accelerating the policy's improvement. Throughout these iterations, the *Q*-function estimates are progressively aligned with actual returns, and the policy is incrementally optimized to better utilize these estimates. The iterative cycle is designed to continue until the policy reaches a convergence point, where it achieves an efficient strategy for the task at hand. Convergence is influenced by the task's complexity, the diversity of experiences, and the tuning of hyperparameters, including the UTD ratio *D* and the ensemble size of *Q*-functions. The process ends when the policy stabilizes, indicating that it has effectively learned to maximize expected returns given the environment's dynamics. The whole process is given in Algorithm 1.

E. Computational Complexity Analysis

1) Computational Complexity for Fully Digital Precoding: The computational complexity for fully digital precoding with analog power allocation involves managing continuous power levels across all antennas. In this scenario, the action space is continuous and represented as $[0, \bar{p}_{\max}]^{M\times(K\times N_d)}$. The REDQ algorithm's complexity in updating each *Q*-function is determined by operations such as matrix multiplications and gradient descent steps. Thus, the computational complexity for fully digital precoding with analog power allocation becomes $\mathcal{O}(N_e \times J \times |\mathcal{B}| \times M \times N_d \times K^2)$, where N_e is the number of training steps.

For fully digital precoding with discrete power allocation and action coding, the action space is initially discrete but transformed into a continuous one through action encoding, simplifying the evaluation process. Despite this transformation, the matrix operations involved in the REDQ algorithm remain similar. The complexity for this scenario is also

Algorithm 1 REDQ Algorithm With the Action Coding Scheme

```
Initialize policy network parameters $\Theta$, $J$ $Q$-functions parameters $\Phi_j, j = 1, \dots, J$, empty experience buffer $\mathcal{O}$. Assign initial target $Q$-function parameters $\Phi_{\mathrm{targ},j}$ to match $\Phi_j$ for each $j$.
for training step $t = 1$ to $N_e$ do
    Execute an action $\mathbf{a}(t)$ chosen according to $\pi_\Theta(\cdot|\mathbf{s}(t))$ and conduct action coding scheme, observe the resulting reward $r(t)$ and next state $\mathbf{s}(t+1)$.
    Store the observed transition $(\mathbf{s}(t), \mathbf{a}(t), r(t), \mathbf{s}(t+1))$ in the experience buffer $\mathcal{O}$ by appending it to the existing contents: $\mathcal{O} \leftarrow \mathcal{O} \cup \{(\mathbf{s}(t), \mathbf{a}(t), r(t), \mathbf{s}(t+1))\}$.
    for $D$ rounds of updates do
        Draw a mini-batch $\mathcal{B}$ of transitions $\{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')\}$ from $\mathcal{O}$.
        Randomly select a subset size $J'$ from the set $\{1, 2, \dots, J\}$ to determine which $Q$-functions to update.
        Calculate the target value $q$ for updating all $J$ $Q$-functions as equation (37).
        for each $Q$-function $j = 1$ to $J$ do
            Adjust $\Phi_j$ by minimizing the squared difference between its current estimates and the target $q$, using gradient descent as (38).
            Update the target network parameters as (39).
        end for
        Update the policy network parameters $\Theta$ to maximize the expected reward as (40).
    end for
end for
```

$\mathcal{O}(N_e \times J \times |\mathcal{B}| \times M \times N_d \times K^2)$, reflecting the computational demands of managing the additional data streams in both analog and discrete settings.

In the case of fully digital precoding with discrete power allocation without action coding, the action space $|\mathcal{P}|^{M \times (K \times N_d)}$ requires evaluating all possible power levels for each antenna, data stream, and user. This results in a significant increase in complexity due to the exponential growth in the number of possible actions. The resulting complexity is $\mathcal{O}(N_e \times J \times |\mathcal{B}| \times |\mathcal{P}|^{M \times (K \times N_d)} \times K)$, where $|\mathcal{P}|$ is the number of discrete power levels.

2) Computational Complexity for Hybrid Precoding: In hybrid precoding with analog power allocation, the action space is continuous and spans $[0, \bar{p}'_{\max}]^M$, reflecting the need to allocate power at each antenna, even though fewer RF chains are used. Each update in the REDQ algorithm involves matrix multiplications that scale with the combined dimensions of K. Therefore, the computational complexity for this scenario is $\mathcal{O}(N_e \times J \times |\mathcal{B}| \times M \times K)$.

For hybrid precoding with discrete power allocation and action coding, the action space, although discrete, is encoded into a continuous one to simplify the update process. The matrix operations still reflect the dimensions of all antennas, data streams, and users, leading to a complexity of $\mathcal{O}(N_e \times J \times |\mathcal{B}| \times M \times K)$.

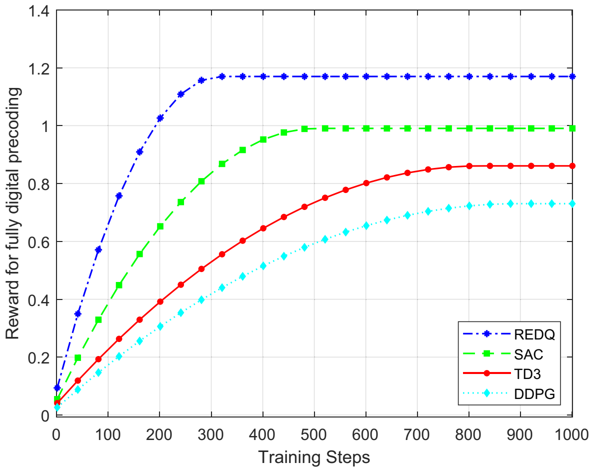

Fig. 2. Comparison of different RL algorithms for fully digital precoding.

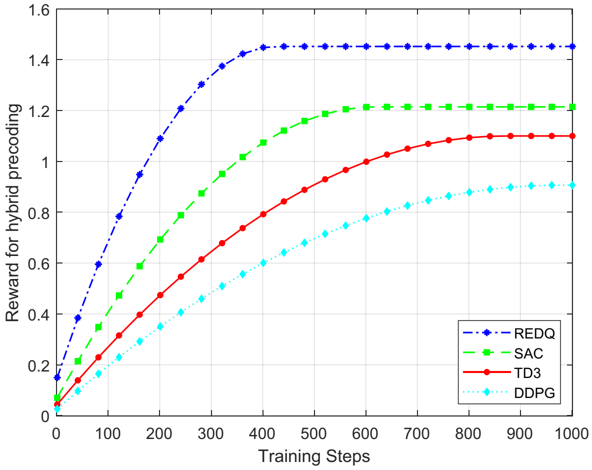

Fig. 3. Comparison of different RL algorithms for hybrid precoding.

In hybrid precoding with discrete power allocation without action coding, the complexity increases significantly due to the need to evaluate a fully discrete action space of $|\mathcal{P}|^M$, where each of the M antennas has discrete power levels to choose from. This results in an exponential increase in computational requirements, as all possible power levels across antennas must be considered. The complexity for this scenario is therefore $\mathcal{O}(N_e \times J \times |\mathcal{B}| \times |\mathcal{P}|^M \times K)$.

VII. Simulation Results

In simulations, we consider a scenario where the number of users is $K = 4$ and the number of antennas of the BS and user are $M = 64$ and $N = 16$. In the hybrid precoding, the number of data streams $N_d = 2$, the penalty parameters $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$. In the discrete power allocation, the number of power levels $|\mathcal{P}| = 3$. We define the upper bound of the total power $P_0 = P'_0 = 30\,\mathrm{dBm} = 1000$ mW. In the reinforcement learning algorithm, the discount factor $\gamma = 0.9$, the batch size is $|\mathcal{B}| = 32$, and the learning rates $\omega = 0.01$ and the number of steps of each episode $N_e = 500$. The number of Q-functions $J = 10$, and the updated Q-function subset size

$J' = 2$. The number of updates $D = 20$. All simulation results are averaged over multiple simulations.

To demonstrate the efficacy of the REDQ algorithm, it is compared with established reinforcement learning algorithms including SAC, TD3, and DDPG as baselines. The comparison, illustrated in Figures 2 and 3, focuses on analog power allocation scenarios incorporating HWI. These results underscore the superior performance of the REDQ algorithm, attributable to its significant enhancements in sample efficiency and its capability to mitigate overestimation bias. By employing an ensemble of Q-functions and selectively updating them with a randomized subset, REDQ markedly diversifies the process of value estimation, yielding more stable and accurate predictions. Furthermore, REDQ's high UTD ratio, achieved via numerous updates per environmental interaction and the judicious use of the replay buffer, ensures a more effective learning process from each data point. This strategy not only addresses the overestimation bias endemic but also expedites the learning process by fully leveraging the available data. Moreover, the algorithm's model-free design minimizes the complexity and potential inaccuracies typical of model-based approaches, enhancing REDQ's robustness and superior performance in complex environments.

In Figure 4, the reward for fully digital precoding, denoted in (35), is depicted under conditions of discrete and analog power allocations (PA), incorporating the presence and absence of HWI. An increase in training steps prompts the rewards associated with all four conditions to converge smoothly, thus demonstrating the efficacy of the REDQ algorithm. It is discernible that the reward involving HWI is less than that without HWI. This disparity arises from introducing additional noise, distortion, and power inefficiencies by HWI into the system. The augmented noise and distortion engendered by HWI elevate the power requisites for achieving the sought-after data rate. Consequently, this leads to decreased energy efficiency, as more power is consumed per unit of information transmitted or received. Moreover, the reward linked to analog power allocation exceeds that of discrete power allocation. In discrete power allocation, each antenna is assigned a specific power level. Analog power allocation allows for real-valued power adjustment, resulting in more efficient power utilization. It also shows that the discrete power allocation model converges faster than the analog power allocation model. This is because discrete power allocation operates within a finite and discrete action space, implying that the agent can select power levels from a restricted set of options. This confines the action space, simplifying the action selection process and mitigating the complexity of the learning problem. However, analog power allocation features a real-valued action space, thus rendering the action selection process more intricate. This demands more exploration and time to identify an efficient power allocation.

Figure 5 depicts the reward for hybrid precoding, as indicated in (36), plotted against the number of training steps. It is crucial to note that the reward for hybrid precoding cannot be directly compared to fully digital precoding, owing to the use of disparate reward functions in each case. It can be discerned that the reward for hybrid precoding also rises, and convergence is seen with the increase in training steps.

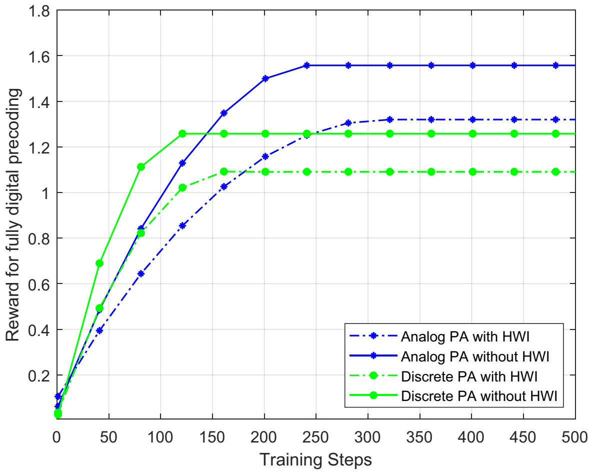

Fig. 4. Average reward of fully digital precoding versus training steps.

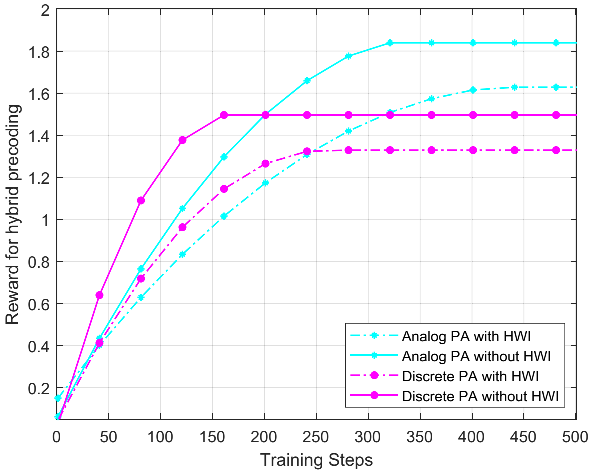

Fig. 5. Average reward of hybrid precoding versus training steps.

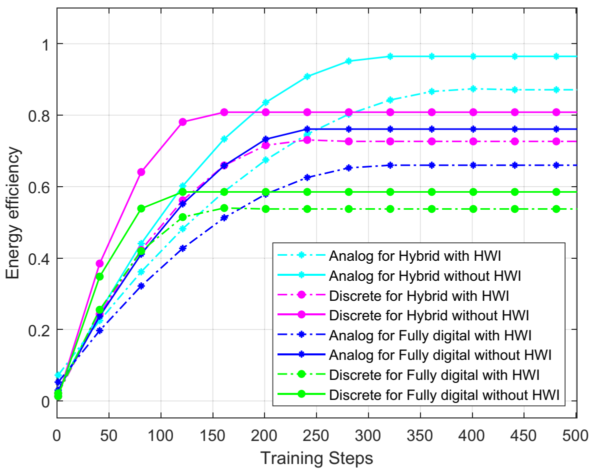

Fig. 6. Energy efficiency versus training steps.

Figure 6 illustrates the energy efficiency, under all conditions as a function of the training steps. In all scenarios, energy efficiency tends to increase and converge smoothly. The energy

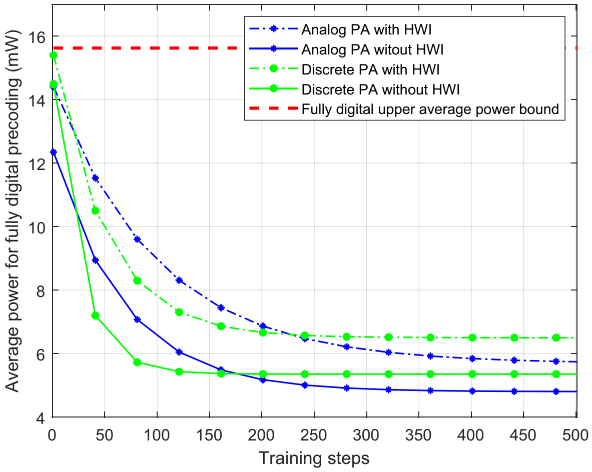

Fig. 7. Average power for fully digital precoding versus training steps.

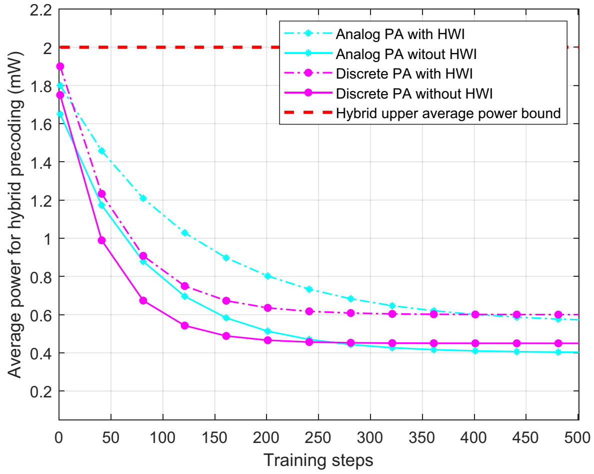

Fig. 8. Average power for hybrid precoding versus training steps.

efficiency corresponding to hybrid precoding surpasses that of fully digital precoding. This can be attributed to hybrid precoding's ability to amalgamate the advantages of both analog and fully digital precoding. Analog precoding can achieve high power efficiency by mapping the transmitted signal to a limited number of RF chains. This reduces the overall power consumption compared to fully digital precoding, where each antenna requires an individual RF chain.

In figure 7, the average power of each antenna in the fully digital precoding is plotted, which is obtained by $\frac{C_{\text{tot}}(\boldsymbol{P})}{M}$. The unit of the power is milliwatt (mW). It shows as the training steps increase, the average power decreases and converges below the upper bound. The upper bound $\bar{p}_{\max} = \frac{P_0}{M}$. Figure 8 plots the average power per antenna, per user, and per data stream in the hybrid precoding, which is obtained by $\frac{C'_{\text{tot}}(\boldsymbol{P}')}{MKN_d}$. The upper bound is $\bar{p}'_{\max}$, which is set as 2 mW. From these figures, we can observe that the average power of the discrete power allocation is higher than analog power allocation. This disparity is because the available power levels of discrete power allocation are either limited or suboptimal. As a result, discrete power allocation leads to higher power consumption than analog power allocation, which is more granular.

Figure 9 shows the total power (presented as $C_{\text{tot}}(\boldsymbol{P})$ for the fully digital precoding and $C'_{\text{tot}}(\boldsymbol{P}')$ for hybrid precoding), which is the sum of the transmitted power of the BS, versus the training steps. Since we set the upper total power bounds for the hybrid and digital models $P_0 = P'_0$, we only need to plot one upper total power bound. This figure shows that the total power associated with both hybrid and fully digital precoding consistently falls within their respective upper and lower total power bounds. In addition, the total power consumed in discrete power allocation surpasses analog power allocation. Furthermore, the total power associated with fully digital precoding exceeds that linked to hybrid precoding. This can be attributed to the efficient shaping of transmitted signals in the analog domain by hybrid precoding, thereby reducing overall power requirements. The graph also indicates that the total power associated with a model incorporating HWI is higher than without HWI. This increase is because HWI introduces additional imperfections and non-idealities in the system's hardware components, leading to elevated power consumption.

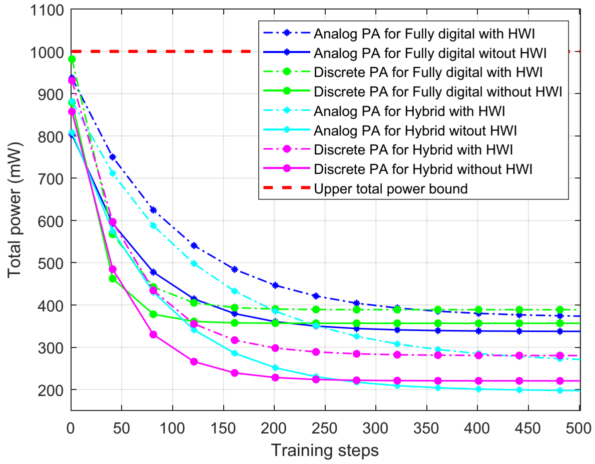

Fig. 9. Total power versus training steps.

Figure 10 presents total power versus the user number K. In this simulation, the total power constraint P_0 and P'_0 increase proportionally with the number of users. For instance, $P_0 = P'_0 = 500$ when $K = 2$, since we defined $P_0 = P'_0 = 1000$ when $K = 4$ previously. From the graph, it is evident that an increase in the number of users necessitates a larger distribution of total power, thus resulting in an elevated total power requirement. Additionally, when the user count is low, adding more users leads to a significant surge in the required total power. However, as the user number expands, the additional power demand for each new user begins to decline. This is attributable to enhanced power allocation efficiency and the effective distribution of available power resources across a larger user pool. Furthermore, the decreasing slope also signifies that the incremental power needed to support additional users diminishes as the population grows. This phenomenon arises due to harnessing multiple antennas, the system can spatially isolate users, thereby mitigating the impact of interference, see [35] for advanced interference mitigation techniques.

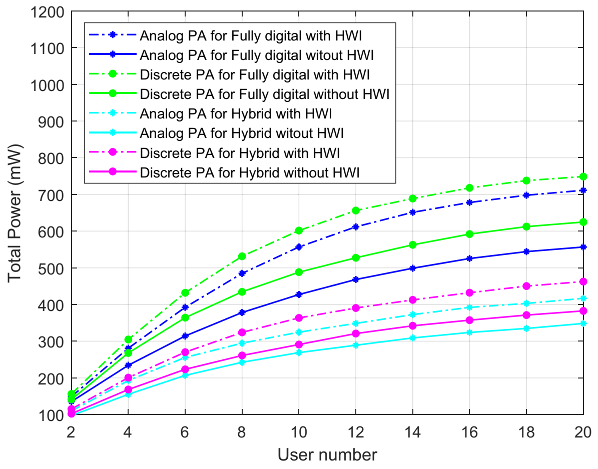

Fig. 10. Total power versus users number.

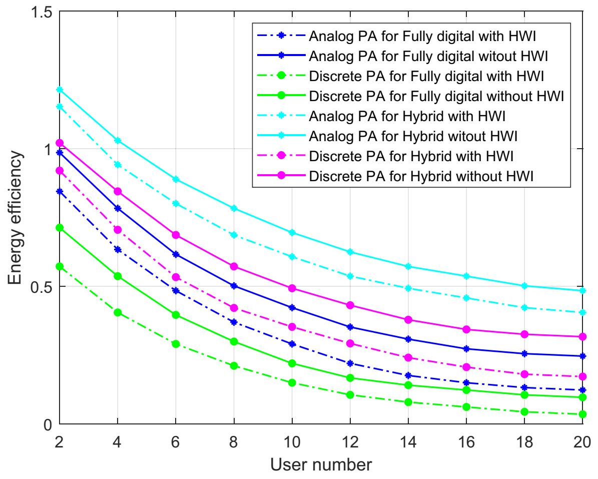

Fig. 11. Energy efficiency versus users number.

TABLE II
COMPARISON OF COMPLEXITIES FOR DIFFERENT TECHNIQUES

Techniques	Complexities	Complexity increase without AC
FD Analog PA	3.28×10^8	
FD Discrete PA with AC	3.28×10^8	
FD Discrete PA without AC	1.24×10^{250}	$\times\ 3.77 \times 10^{241}$
HB Analog PA	4.09×10^7	
HB Discrete PA with AC	4.09×10^7	
HB Discrete PA without AC	1.10×10^{36}	$\times\ 2.68 \times 10^{28}$

Figure 11 demonstrates the energy efficiency versus the user number K. As the number of users in a communication system increases, energy efficiency tends to decrease. This is because, with more users, the system must allocate power across a larger set of channels to maintain the desired QoS, leading to an increase in the total power required for transmission. Additionally, increasing the number of users exacerbates the complexity of the precoding process of zero-forcing transmission. To eliminate inter-user interference, the precoder must work harder to orthogonalize the signals intended for different users, leading to less power-efficient transmission strategies.

Table II compares the computational complexities of various PA schemes used in fully digital (FD) and hybrid (HB) precoding systems with and without action coding (AC). The complexities are measured in terms of the number of operations, which represents the total number of computational operations required, using parameters $N_e = 500$, $J = 10$, $|\mathcal{B}| = 32$, $|\mathcal{P}| = 3$, $N_d = 2$, $K = 4$, $M = 64$ and formulas derived in the above section. The complexity increase without AC represents how much the complexity of the PA scheme increases without the AC method. This column highlights the dramatic rise in complexity when transitioning from scenarios with action coding to those without it. The value 3.77×10^{241} is derived by comparing the complexity of FD Discrete PA without AC, 1.24×10^{250}, to FD Discrete PA with AC, 3.28×10^8. This ratio shows the extent of computational increases without implementing AC. Similarly, the value 2.68×10^{28} is calculated by comparing the complexity of HB Discrete PA without AC, 1.10×10^{36}, to HB Discrete PA with AC, 4.09×10^7. These comparisons highlight the significant efficiency gains by the AC method in reducing the computational burden of discrete PA processes.

VIII. CONCLUSION

This paper proposes analog and discrete power allocation under dynamic QoS requirements to maximize the long-term energy efficiency in mMIMO systems with and without HWI. We investigated both fully digital and hybrid precoding techniques. Per antenna and total power constraints are applied to the problem. For this constrained long-term optimization problem, the REDQ algorithm is modified with a proposed action coding scheme, which makes the discrete power allocation for numerous antennas possible. Extensive simulations are conducted to demonstrate the effectiveness of the proposed power allocation strategies across different scenarios. Future research will extend the current work by incorporating imperfect CSI conditions. This includes considering CSI estimation errors, feedback delays, and robust optimization techniques to handle these imperfections. Addressing imperfect CSI will ensure that the proposed algorithm remains effective under more realistic and practical scenarios, further enhancing the robustness and applicability of our approach.

REFERENCES

[1] N. Garg, M. Sellathurai, V. Bhatia, and T. Ratnarajah, "Function approximation based reinforcement learning for edge caching in massive MIMO networks," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2304–2316, Apr. 2021.

[2] Z. Liu, N. Garg, and T. Ratnarajah, "Multi-agent federated reinforcement learning strategy for mobile virtual reality delivery networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 1, pp. 100–114, Jan./Feb. 2024.

[3] S. Biswas, C. Masouros, and T. Ratnarajah, "Performance analysis of large multiuser MIMO systems with space-constrained 2-D antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3492–3505, May 2016.

[4] A. K. Papazafeiropoulos and T. Ratnarajah, "Deterministic equivalent performance analysis of time-varying massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5795–5809, Oct. 2015.

[5] Z. Liu, N. Garg, and T. Ratnarajah, "Multi-agent federated *Q*-learning algorithms for wireless edge caching," *IEEE Trans. Veh. Technol.*, early access, Oct. 2, 2024, doi: 10.1109/TVT.2024.3473738.

[6] Z. Liu, H. Song, and D. Pan, "Distributed video content caching policy with deep learning approaches for D2D communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15644–15655, Dec. 2020.

[7] C. B. Papadias, T. Ratnarajah, and D. T. Slock, *Spectrum Sharing: The Next Frontier in Wireless Networks*. Hoboken, NJ, USA: Wiley, 2020.

[8] A. Nazábal and J. Vía, "Analog antenna combining in multiuser OFDM systems: Beamforming design and power allocation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2664–2667.

[9] H.-H. Nguyen and W.-J. Hwang, "Distributed scheduling and discrete power control for energy efficiency in multi-cell networks," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2198–2201, Dec. 2015.

[10] M. Olyaee, M. Eslami, and J. Haghighat, "An energy-efficient joint antenna and user selection algorithm for multi-user massive MIMO downlink," *IET Commun.*, vol. 12, no. 3, pp. 255–260, 2018.

[11] A. Konar and N. D. Sidiropoulos, "A simple and effective approach for transmit antenna selection in multiuser massive MIMO leveraging submodularity," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4869–4883, Sep. 2018.

[12] H. Li, J. Cheng, Z. Wang, and H. Wang, "Joint antenna selection and power allocation for an energy-efficient massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 257–260, Feb. 2019.

[13] D. Park, "Sum rate maximisation with transmit antenna selection in massive MIMO broadcast channels," *Electron. Lett.*, vol. 54, no. 21, pp. 1245–1247, 2018.

[14] Y. He et al., "Energy efficient power allocation for cell-free mmWave massive MIMO with hybrid precoder," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 394–398, Feb. 2022.

[15] A. Koc and T. Le-Ngoc, "Swarm intelligence based power allocation in hybrid millimeter-wave massive MIMO systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–7.

[16] A. Koc, M. Wang, and T. Le-Ngoc, "Deep learning based multi-user power allocation and hybrid precoding in massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 5487–5492.

[17] X. Su and Y. Jiang, "Optimal zero-forcing hybrid downlink precoding for sum-rate maximization," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 463–467, Mar. 2022.

[18] A. J. Ortega, R. Sampaio-Neto, and R. P. David, "Hybrid precoding design and power allocation for massive MU-MIMO mmWave systems," in *Proc. Int. Wireless Commun. Mobile Comput.*, 2022, pp. 50–55.

[19] V. Radhakrishnan, O. Taghizadeh, and R. Mathar, "Energy efficient full duplex massive MIMO multi-carrier bidirectional communication with hardware impairments," in *Proc. Int. ITG Workshop Smart Antennas*, 2019, pp. 1–8.

[20] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[21] L. Van der Perre, L. Liu, and E. G. Larsson, "Efficient DSP and circuit architectures for massive MIMO: State of the art and future directions," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4717–4736, Sep. 2018.

[22] J. Zhang, H. Luo, N. Garg, A. Bishnu, M. Holm, and T. Ratnarajah, "Design and analysis of wideband in-band-full-duplex FR2-IAB networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4183–4196, Jun. 2022.

[23] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[24] S. Dankwa and W. Zheng, "Twin-delayed DDPG: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proc. 3rd Int. Conf. Vis. Image Signal Process.*, 2019, pp. 1–5.

[25] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[26] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4935–4945, Jun. 2019.

[27] F. Sangare, D. H. N. Nguyen, and Z. Han, "Learning frameworks for dynamic joint RF energy harvesting and channel access," *IEEE Access*, vol. 7, pp. 84524–84535, 2019.

[28] N. Garg, M. Sellathurai, and T. Ratnarajah, "Reinforcement learning based per-antenna discrete power control for massive MIMO systems," in *Proc. 54th Asilomar Conf. Signals, Syst. Comput.*, 2020, pp. 1028–1032.

[29] J. Zhang, N. Garg, and T. Ratnarajah, "Design of in-band-full-duplex IAB networks for integrated sensing and communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 1888–1893.

[30] A. Khan, S. Wang, and Z. Zhu, "Angle-of-arrival estimation using an adaptive machine learning framework," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 294–297, Feb. 2019.

[31] A. M. Elbir, "DeepMUSIC: Multiple signal classification via deep learning," *IEEE Sens. Lett.*, vol. 4, no. 4, pp. 1–4, Apr. 2020.

[32] X. Chen, C. Wang, Z. Zhou, and K. Ross, "Randomized ensembled double Q-learning: Learning fast without a model," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–25.

[33] N. Garg and G. Sharma, "Analog precoder feedback schemes with interference alignment," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5382–5396, Aug. 2018.

[34] N. Garg, M. Sellathurai, and T. Ratnarajah, "In-network caching for hybrid satellite-terrestrial networks using deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 8797–8801.

[35] H. Du, T. Ratnarajah, M. Sellathurai, and C. B. Papadias, "Reweighted nuclear norm approach for interference alignment," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3754–3765, Sep. 2013.

Zhikai Liu received the B.S. degree in communications engineering from Hohai University in 2014, and the M.S. degree in electromagnetic field and microwave technology from South China Normal University in 2021. He is currently pursuing the Ph.D. degree with The University of Edinburgh. His research interests include edge caching, distributed learning, signal processing, and wireless communications.

Navneet Garg (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the College of Science and Engineering, Jhansi, India, in 2010, the M.Tech. degree in digital communications from the ABV-Indian Institute of Information Technology and Management, Gwalior, in 2012, and the Ph.D. degree from the Department of Electrical Engineering, Indian Institute of Technology Kanpur, India, in June 2018. From July 2018 to January 2019, he visited The University of Edinburgh, U.K. Next, he was employed as a Postdoctoral Research Associate for almost 4.5 years, that is, from February 2019 to 2020 with Heriot-Watt University, Edinburgh; from February 2020 to 2023 with The University of Edinburgh; and from April to July 2023 with the Indian Institute of Technology Indore, India. Since August 2023, he has been working as an Assistant Professor with The LNM Institute of Information Technology, Jaipur, India. His main research interests include wireless communications, signal processing, optimization, and machine learning.

Tharmalingam Ratnarajah (Senior Member, IEEE) works as a Fred Harris Endowed Chair of Digital Signal Processing with San Diego State University, and the Director of the Communication Systems and Signal Processing Institute. He was the Lead Coordinator of the European Union Projects HARP (4.6M€) in the area of highly distributed MIMO and ADEL (3.7M€) in the area of licensed shared access. He was also the Coordinator of the European Union Future and Emerging Technologies Project CROWN (3.4M€) in the area of cognitive radio networks and HIATUS (3.6M€) in the area of interference alignment. His research interests include signal processing and information-theoretic aspects of beyond 5G cellular networks, full-duplex radio, mmWave communications, random matrix theory, big data analytics and machine learning for wireless networks, statistical and array signal processing, physical-layer secrecy, and interference alignment. He has published over 475 Peer-reviewed papers in these areas and holds four U.S. patents. He was an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2015 to 2017, and the Technical Co-Chair of the 17th IEEE International Workshop on Signal Processing Advances in Wireless Communications, Edinburgh, U.K., 3–6 July 2016. He is a Fellow of the Higher Education Academy.