Pervasive Machine Learning for Smart Radio Environments Enabled by Reconfigurable Intelligent Surfaces

By George C. Alexandropoulos🄳 , *Senior Member IEEE*,
Kyriakos Stylianopoulos🄳 , *Graduate Student Member IEEE*, Chongwen Huang🄳 , *Member IEEE*,
Chau Yuen🄳 , *Fellow IEEE*, Mehdi Bennis🄳 , *Fellow IEEE*, and Mérouane Debbah, *Fellow IEEE*

George C. Alexandropoulos is with the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15784 Athens, Greece, and also with the Technology Innovation Institute, Abu Dhabi, United Arab Emirates (e-mail: alexandg@di.uoa.gr).

Kyriakos Stylianopoulos is with the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15784 Athens, Greece (e-mail: kstylianop@di.uoa.gr).

Chongwen Huang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, with the International Joint Innovation Center, Zhejiang University, Haining 314400, China, and also with the Zhejiang Provincial Key Laboratory of Information Processing, Communications and Networks (IPCAN), Hangzhou 310027, China (e-mail: chongwenhuang@zju.edu.cn).

Chau Yuen is with the Engineering Product Development (EPD) Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: yuenchau@sutd.edu.sg).

Mehdi Bennis is with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland (e-mail: mehdi.bennis@oulu.fi).

Mérouane Debbah is with the Technology Innovation Institute, Abu Dhabi, United Arab Emirates, and also with the CentraleSupélec, University Paris-Saclay, 91192 Gif-sur-Yvette, France (e-mail: merouane.debbah@tii.ae).

Digital Object Identifier 10.1109/JPROC.2022.3174030

ABSTRACT | The emerging technology of reconfigurable intelligent surfaces (RISs) is provisioned as an enabler of smart wireless environments, offering a highly scalable, low-cost, hardware-efficient, and almost energy-neutral solution for dynamic control of the propagation of electromagnetic signals over the wireless medium, ultimately providing increased environmental intelligence for diverse operation objectives. One of the major challenges with the envisioned dense deployment of RISs in such reconfigurable radio environments is the efficient configuration of multiple metasurfaces with limited, or even the absence of, computing hardware. In this article, we consider multiuser and multi-RIS-empowered wireless systems and present a thorough survey of the online machine learning approaches for the orchestration of their various tunable components. Focusing on the sum-rate maximization as a representative design objective, we present a comprehensive problem formulation based on deep reinforcement learning (DRL). We detail the correspondences among the parameters of the wireless system and the DRL terminology, and devise generic algorithmic steps for the artificial neural network training and deployment while discussing their implementation details. Further practical considerations for multi-RIS-empowered wireless communications in the sixth-generation (6G) era are presented along with some key open research challenges. Different from the DRL-based status quo, we leverage the independence between the configuration of the system design parameters and the future states of the

wireless environment, and present efficient multiarmed bandits approaches, whose resulting sum-rate performances are numerically shown to outperform random configurations, while being sufficiently close to the conventional deep Q network (DQN) algorithm, but with lower implementation complexity.

KEYWORDS | Artificial neural networks (ANNs); deep reinforcement learning (DRL); future wireless networks; reconfigurable intelligent surface (RIS); smart radio environment.

NOMENCLATURE

A. Notations Used for Wireless Communications

$\mathbb{C}$ and $\mathbb{R}$	Complex and real number sets, respectively.
K	Number of single-antenna UEs; UE_k denotes the kth UE with $k = 1, 2, \ldots, K$.
N_T	Number of BS antenna elements.
M	Number of identical RISs; RIS_m denotes the mth RIS with $m = 1, 2, \ldots, M$.
$N \triangleq N_h N_v$	Number of unit elements per RIS with $i = 1, 2, \ldots, N$: N_h placed in the horizontal and N_v in the vertical dimensions.
$N_{tot} \triangleq MN$	Total number of phase-tunable elements of all RISs.
N_{group}	Number of RIS elements whose phase configurations are set together to the same value.
$N_{control}$	Total number of individually controllable elements from all reconfigurable intelligent surfaces (RIS).
$\mathbf{H}_m$	$N \times N_T$ channel matrix for the RIS_m-BS link.
$\mathbf{g}_{k,m}$	N-element row vector for the UE_k-RIS_m channel.
$\mathbf{h}_k$	N_T-element row vector for the UE_k-BS channel.
$L(d)$	Pathloss attenuation at a certain distance d; d_k, d_m, and $d_{k,m}$ denote the geometrical distances of the UE_k-BS, RIS_m-BS, and UE_k-RIS_m links, respectively.
κ_1 and κ_2	Ricean factors of each RIS_m-BS and each UE_k-RIS_m link, respectively.
$\mathbf{f}_u(\varphi, \vartheta)$	Array response row vector for a multiantenna/multielement node $u \in \{BS, RIS_m\}$ with azimuth and elevation angles of arrival/departure φ and ϑ, respectively.
λ	Signal wavelength.
d_{RIS} and d_{BS}	Spacing of adjacent meta-atoms at each RIS_m and that of adjacent antenna elements at the BS.
b	Phase resolution in bits of each RIS element.
ϕ_m and $\mathbf{\Phi}_m$	N-element phase profile vector for RIS_m and $\mathbf{\Phi}_m \triangleq \text{diag}\{\phi_m\}$.
$\mathcal{F}$	Set of discrete phase states at each RIS, i.e., $[\phi_m]_i \in \mathcal{F} \, \forall m, i$.
P	Total power budget at the BS for transmission.
$\mathbf{x} \triangleq \mathbf{Vq}$	N_T-element transmitted signal vector from BS composed by the $N_T \times K$ precoding matrix $\mathbf{V}$ and the K-element symbol vector $\mathbf{q} \triangleq [q_1, q_2, \ldots, q_K]^T$ with $\mathbb{E}\{\mathbf{qq}^H\} = \frac{P}{K}\mathbf{I}_K$.
$\mathcal{V}$	Set of available $N_T \times 1$ BS precoding vectors of unit norm for all UEs, i.e., $[\mathbf{V}]_{:,k} \in \mathcal{V} \, \forall k$.
y_k	Received signal at UE_k in the baseband.
ν	Attenuation factor for the power of all direct UE_k-BS links such that $\nu \in [0, 1]$.
$n_k \sim \mathcal{CN}(0, \sigma^2)$	Additive white Gaussian noise (AWGN) at each received signal with zero mean and variance σ^2.
f	Carrier frequency.
$\tilde{R}_k$	Received data rate by user k.
R_k^{req}	Requested data rate by user k.

B. Notations Used for Machine Learning

t	Discrete time step.
$\mathbf{s}_t \in \mathcal{S}$	MDP state at time t chosen from the available state set $\mathcal{S}$.
$\dim(\mathbf{s}_t)$	Dimensionality of an MDP state vector $\mathbf{s}_t$.
$\mathbf{a}, \mathbf{a}_t \in \mathcal{A}$	MDP actions chosen from the available action space set $\mathcal{A}$ with cardinality $\text{card}(\mathcal{A})$; $\mathbf{a}_t$ denotes the action chosen at time t.
$\Pr[\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t]$	MDP transition probability: observing s_{t+1} while being at s_t and acting with $\mathbf{a}_t$.
$r_t \triangleq \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t)$	Reward at time t via the reward function.
S_t, A_t, and R_t	Random variables for the state, action, and reward at time t.
T	Final time step of a finite MDP.
$\varpi(\mathbf{a}_t|\mathbf{s}_t)$	Agent's policy at time t, i.e., the likelihood of selecting action $\mathbf{a}_t$ when being at $\mathbf{s}_t$.
G_t^ϖ	Return at time t after using policy $\varpi(\mathbf{a}_t|\mathbf{s}_t)$.
γ	Discount factor such that $\gamma \in (0, 1]$.
$V^\varpi(\mathbf{s})$	Value function of policy for any $\mathbf{s} \in \mathcal{S}$.
$Q^\varpi(\mathbf{s}, \mathbf{a})$	Action value function of policy for any $\mathbf{s} \in \mathcal{S}$ and any action $\mathbf{a} \in \mathcal{A}$.
$V^*(\mathbf{s})$, $Q^*(\mathbf{s}, \mathbf{a})$	Optimal value and action value functions.
$\mathbf{w}$ and $\mathbf{w}^-$	Vectors with the weight parameters of neural networks.
$Q_\mathbf{w}$ and $\bar{Q}_{\mathbf{w}^-}$	Neural network and a target neural network used by DQN.

τ	"Temperature" of the soft-update of the weights of the target network used by DQN.
$\hat{G}_{\mathbf{w}}(\mathbf{s})$	Reward-prediction neural network, with weights $\mathbf{w}$, used by the neural ϵ-greedy algorithm, having an observation $\mathbf{s}$ as input.
$J(\cdot)$	Abstract objective function of a general DRL algorithm.
$\mathcal{D}$	Batch of experience tuples.
t'	Update interval during training of a DRL algorithm.
$\mathcal{L}(\mathbf{w})$	Objective function of a DQN.
$\hat{\mathcal{L}}(\mathbf{w})$	Loss function of the neural ϵ-greedy algorithm.
α	Learning rate of gradient descent or ascent.
ϵ	Probability of selecting a random action.
$\mathcal{U}(\mathcal{A})$	Selection operation from the discrete set $\mathcal{A}$ with probability $1/\mathrm{card}(\mathcal{A})$.
$J_{\mathrm{PG}}(\mathbf{w})$	Objective function of a policy gradient method.
τ	Trajectory of an MDP episode.
$\mathrm{A}(\mathbf{s}_t, \mathbf{a}_t)$	Advantage value function at state $\mathbf{s}_t$ using action $\mathbf{a}_t$.
$\mathrm{G}_t(\mathbf{a})$	Running average of achieved reward for action $\mathbf{a}$ at time t.
$\mathrm{N}_t(\mathbf{a})$	Number of times action $\mathbf{a}$ has been selected at time t.
$r_i(\mathbf{a})$	Reward when action a is selected for the ith time.
c	Width of the confidence interval of the UCB strategy.
$\mathbb{I}_{\mathcal{A}}(\mathbf{a})$	Enumeration operator that maps a discrete action $\mathbf{a}$ to its index in the discrete action space $\mathcal{A}$.

C. Abbreviations of This Article

5G	Fifth generation.
6G	Sixth generation.
ANN	Artificial neural network.
AoA	Angle of Arrival.
AoD	Angle of Departure.
AoI	Age of Information.
AWGN	Additive white Gaussian noise.
BS	Base station.
CE	Channel estimation.
CSI	Channel state information.
D^{3}QN	Dueling double deep Q network.
DDPG	Deep deterministic policy gradient.
DFT	Discrete Fourier transform.
DQN	Deep Q network.
DRL	Deep reinforcement learning.
EE	Energy efficiency.
EM	Electromagnetic.
eMBB	enhanced Mobile Broadband.

FET	Field-effect transistor.
GPU	Graphical processing unit.
i.i.d.	Independent identically distributed.
IoT	Internet of Things.
LOS	Line-of-sight.
LSTM	Long short-term memory.
MAC	Medium access control.
MAML	Multiagent reinforcement learning.
MEMS	Microelectromechanical system.
MDP	Markov decision process.
MIMO	Multiple-input multiple-output.
MISO	Multiple-input single-output.
mmWave	Millimeter wave.
MRT	Maximum ratio transmission.
mse	Mean square error.
NLOS	Non-line-of-sight.
NMSE	Normalized mean square error.
NOMA	Nonorthogonal multiple access.
NR	New radio.
OFDM	Orthogonal frequency-division multiplexing.
PG	Policy gradient.
POMDP	Partially observable Markov decision process.
PPO	Proximal policy optimization.
QoS	Quality of Service.
RF	Radio frequency.
RIS	Reconfigurable intelligent surface.
RL	Reinforcement learning.
SINR	Signal-to-interference-plus-noise ratio.
SNR	Signal-to-noise ratio.
TD	Temporal difference.
THz	Terahertz.
TTI	Transmission time interval.
UAV	Unmanned autonomous vehicle.
UCB	Upper confidence bound.
UE	User equipment.
URLLC	Ultrareliable and low latency communication.

I. INTRODUCTION

While the Third Generation Partnership Project (3GPP) is finalizing the Release 17 [1] with enhancements on the 5G NR, and the major telecommunications operators are deploying 5G wireless networks around the world [2], academia and industry in wireless communications are already focusing on the Release 18 [3] (termed 5G-Advanced) and working on the definition and identification of requirements and candidate technologies for the 6G of wireless communications [4]. As per the latest consensus [5], [6], 6G networks will require, among other metrics, higher data rates reaching up to 1-Tb/s peak values with extended coverage, 1000× network capacity, and 10× energy and cost efficiency compared to 5G, cm-level positioning accuracy, and $10^8/\mathrm{km}^3$

density of wireless connections. To achieve the envisioned requirements enabling various immersive (e.g., virtual and augmented reality), massive connectivity [7] (e.g., IoT) in industry and smart cities), URLLC (e.g., autonomous vehicles and remote surgery), and zero-touch [8] (i.e., increased network automation) applications, 6G networks need to transform to a unified communication, sensing, and computing platform incorporating artificial intelligence and machine learning methodologies [9]. This platform will highly likely integrate spectra from higher than the 5G mmWave band and the THz frequencies [10], requiring cost- and power-efficient RF front ends and multiantenna transceiver architectures, as well as ultra-wideband waveforms and computationally efficient signal processing schemes [11].

The potential of RISs for programmable EM wave propagation has recently motivated extensive academic and industrial interests, as an enabler for smart wireless environments in the 6G era [4], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. The RIS technology, which typically refers to artificial planar structures with almost passive electronic circuitry (i.e., without any power amplification), is envisioned to be jointly optimized with conventional wireless transceivers [14], [22] in order to significantly boost wireless communications in terms of coverage, spectral and energy efficiency, reliability, and security while satisfying regulated EM field emissions. The envisioned dense deployment of RISs over entities of the wireless environment, other than the actual network devices, is considered a revolutionary means to transform them into network entities with reconfigurable properties, providing increased environmental intelligence for diverse communication objectives [19]. The typical unit element of an RIS is the meta-atom, which is usually fabricated to realize multiple discrete phase states corresponding to distinct EM responses. By externally controlling the states of the meta-atoms in such metasurfaces, various reflection and scattering profiles can be emulated [17]. The RIS technology up to date mainly includes meta-atoms of ultralow power consumption for tuning [18], in the sense that an RIS does not include any power amplifying circuitry. Such metasurfaces can only act as tunable reflectors and, thus, neither receive nor transmit on their own. While almost passive RISs can enable smart radio propagation environments, their purely reflective operation in conjunction with their very limited computing and storage capabilities induce notable network design challenges.

One of the major challenges with RIS-empowered smart wireless environments, including multiple BSs and multiple UEs, is CE, which most commonly constitutes a prerequisite for the optimized orchestration of the phase responses of the unit elements (i.e., meta-atoms) of the multiple RISs. CE involves the estimation of multiple channels simultaneously: the direct channels between each BS and each UE, the channels between each RIS and BS, and the channels between each RIS and each UE [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47]. This task becomes more complex when the deployed RISs are equipped with large numbers of unit elements having nonlinear hardware characteristics [38]. For example, in [39], a general framework for cascaded CE in RIS-assisted MIMO systems was introduced by leveraging combined bilinear sparse matrix factorization and matrix completion. A control protocol enabling linear square estimation for the involved channel vectors in the MISO case was presented in [40]. Jensen and De Carvalho [41] designed an optimal CE scheme, where the RIS unit elements follow an optimal series of activation patterns. Kang [42] presented the joint optimal training sequence and reflection pattern to minimize the mse of CE for RIS-aided wireless systems by recasting the corresponding problem into a convex semidefinite programming one. In [43], an RIS-based activity detection and a CE problem were formulated as a sparse matrix factorization, matrix completion, and multiple measurement vector problem that was solved via an approximate message passing algorithm. In [44], a least-squares Khatri–Rao factorization algorithm was presented for RIS-assisted MIMO systems. A transmission protocol for CE and RIS phase profile optimization was proposed in [45] for RIS-enhanced OFDM systems. A holographic version of an RIS was designed in [46], and its application to THz massive MIMO systems was investigated, along with a closed-loop CE scheme. Very recently, in [47], a channel tracking scheme for the uplink of RIS-enabled multiuser MIMO systems was presented, capitalizing on a tensor representation of the received signal and its parallel factor analysis. To avoid the inevitably large overhead of the estimation of high-dimensional channel matrices in RIS-empowered smart wireless environments, more efficient approaches based on fast RIS phase profile management need to be investigated.

Despite the inevitable challenges for the efficient orchestration of all dynamically reconfigurable entities in smart wireless environments, several applications of RIS-empowered wireless communication have been lately considered. Among them belong the adoption of RISs for extending signal coverage [48], [49], [50], [51], [52], [53], enabling or boosting accurate localization [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], and enabling physical-layer security [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]. In [77], the coexistence of eMBB and URLLC services in a cellular network that is assisted by an RIS was studied. A passive beamformer that can achieve the asymptotic optimal performance by controlling the properties of the incident wave at the RIS was designed in [78], under a limited RIS control link and practical reflection coefficients. An asymptotic closed-form expression for the mutual information of a multiantenna transmitter–receiver pair in the presence of multiple RISs, in the large-antenna limit, was presented in [79]. The fundamental capacity limits of RIS-assisted multiuser wireless communication systems were also investigated in [80]. You *et al.* [81] considered

the application of RISs to assist the uplink transmission from multiple UEs to a multiantenna BS and devised an optimization framework for jointly designing the transmit covariance matrices and the RIS phase profile with partial CSI. In [82], active RISs were considered, which can adapt the phase and amplify the magnitude of the reflected incident signal simultaneously with the support of an additional power source. Joint optimization of the RIS phase profile and the BS beamforming vector was presented in [82] with the objective to minimize the BS transmit power. In [83], the problem of energy efficiency optimization for a wireless communication system with distributed RISs was investigated. RIS-empowered device-to-device (D2D) communications underlaying a cellular network were considered in [84], in which an RIS was employed to enhance the desired signals and suppress interference between paired D2D and cellular links.

With the increasing capabilities of computing infrastructure and the wide availability of testbed datasets, machine learning has been established as a prominent tool for pattern recognition. Particularly, the advent of deep learning [85] has enabled ANNs to detect innate structures over training data and infer underlying models without relying on domain knowledge. Accordingly, such data-driven methods have been implemented as alternatives to model-based techniques in various domains of wireless communications with interesting results [8], [9], effectively shaping the road ahead for 6G smart wireless environments [86]. For instance, in [87] and [88], ANNs were trained to predict the positions of UEs based on channel observations under a localization framework. In [89], the same problem was treated both as a classification and a regression task, and a more robust methodology was proposed. In a similar manner, the estimation of large MIMO channels was facilitated by networks trained on pilot signals in [90], [91]. Arguably, in the most common scenario in the literature, deep learning methods have been tasked with learning optimal digital, analog, or hybrid digital/analog beamforming practices [92], [93], [94], [95].

Very recently, researchers from both academia and industry have exhibited a keen interest in applying deep learning to the design of RIS-empowered smart wireless environments. In the light of the large numbers of free parameters to configure, usually introduced by the phase-tunable RIS meta-atoms, it is expected that there are considerable performance improvements to be exploited with the help of trained ANNs compared to conventional nonconvex optimization approaches that are usually iterative, complex, and without convergence guarantees. In particular, deep learning has already been successfully assessed over a number of pertinent scopes for smart radio environments, such as rate maximization through analog beamforming [96], [97], [98], [99], power maximization [100], compressive sensing estimation [101], channel estimation with possibly hybrid reflecting and sensing RISs [96], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], secrecy maximization [112],

and even real-time imaging [113]. Most works focus on applications on the physical layer, whereas others such as [97] and [114] are intended for deploying deep-learning-based systems on the MAC layer. Very recently, Wang *et al.* [115] presented a metalearning approach to jointly design the multiantenna receiver and configure the RIS reflection coefficients in the uplink of RIS-aided multi-user MIMO systems. Focusing on dynamic rich-scattering conditions in [116], a framework for training a deep neural network as a surrogate forward model, to capture the stochastic dependence of wireless channels on the RIS configuration, was devised.

The vast majority of the deep-learning-based orchestration methods for RIS-empowered smart wireless environments adhere to the paradigm of supervised learning. Accordingly, a data collection stage must be introduced prior to the deployment of the orchestration system in order for the model to be sufficiently trained. It is clear that it is often inconvenient to assume a separate training phase for the deployed machine-learning-based systems for various practical wireless communications applications. Not only this has the effect of greater deployment delays but also the trained orchestration algorithm is probably less equipped to deal with potential changes in the environment, resulting in a decoherence between the training data and the ones observed during the deployment phase. Alternatively, in this article, we opt to focus on the methodology of RL, which poses the design objective at hand as an online optimization problem, by training a model through continuous interactions with the environment. Orchestration methods based on RL algorithms are arguably better suited to a wide variety of applications in wireless networks and, consequently, in RIS-empowered smart radio environments. First and foremost, they have the ability to adapt and change their behavior patterns in nonstationary settings. In addition, the data collection process is embedded in the underlying RL algorithms that are designed for more efficient exploration of the parameter space, as the training evolves. Motivated by the strengths of RL for wireless communications and the relevant lately increasing research interest, in this survey paper, we thoroughly overview the latest advances in DRL approaches for smart radio environments enabled by a plurality of RISs. The main contributions of this article are given as follows.

1) We present a comprehensive modeling of smart wireless environments comprising a multiantenna BS, multiple single-antenna UEs, and multiple commonly accessible RISs that are all controlled by the same dedicated network entity (i.e., the RIS orchestration controller).

2) A detailed introduction to the RL theory is provided with the purpose of explaining the principles behind the most prominent DRL algorithms currently in use by the increasingly crowded community working on RIS-aided wireless communications, supplemented by

a thorough taxonomy that emphasizes their different characteristics.

3) A principled application of DRL to RIS-empowered smart radio environments is presented, detailing the correspondences among the design parameters of the wireless system and the DRL terminology. We consider a general sum-rate maximization problem with individual Quality-of-Service (QoS) constraints by the UEs, and a generic RL algorithmic approach is described as a guideline for concrete implementations.

4) An elaborate literature overview of applications of DRL methods in RIS-based wireless systems is presented, with the purpose of discussing common algorithmic practices and use cases. An accompanying discussion focuses on practical aspects, extensions of the provided DRL-based formulation, and key open research challenges.

5) Guided by the careful examination of the theory of RL and the particularities of the RIS-empowered smart wireless environments, an alternative methodology is proposed for the sum-rate maximization objective. The problem is cast as a multiarmed bandits setting, which is a simpler version of the commonly considered Markovian-based formulation.

6) An extensive numerical investigation of the proposed and benchmark (D)RL algorithms, including the optimal policy, is conducted both in the absence and in the presence of individual QoS constraints, as well as for either the full or partial CSI availability cases.

The reminder of this article is organized as follows. Section II introduces the considered RIS-empowered smart wireless environment, channel models, and design optimization formulation, while Section III includes a thorough overview of the theory behind key RL approaches for wireless communications. The incorporation of DRL in RIS-empowered wireless settings is presented in Section IV, including a detailed DRL-based formulation for the general orchestration problem under investigation and a comprehensive overview of the pertinent literature. Practical considerations, design challenges, and important research directions are discussed in Section V. Section VI presents our numerical results for the proposed (D)RL schemes in comparison with benchmark approaches. Finally, Section VII concludes this article.

Notations: Vectors and matrices are denoted by boldface lowercase and boldface capital letters, respectively. The vectorization, transpose, and Hermitian transpose of $\mathbf{A}$ are denoted by $\text{vec}(\mathbf{A})$, $\mathbf{A}^T$, and $\mathbf{A}^H$, respectively, while $\mathbf{I}_n$ ($n \geq 2$) is the $n \times n$ identity matrix, and $\mathbf{0}_n$ is an n-element column vector with zeros. $[\mathbf{A}]_{i,j}$ is the (i,j)th element of $\mathbf{A}$, $[\mathbf{a}]_i$ and $\|\mathbf{a}\|$ denote $\mathbf{a}$'s ith element and Euclidean norm, respectively, and $\text{diag}\{\mathbf{a}\}$ represents a square diagonal matrix with $\mathbf{a}$'s elements in its main diagonal. $|a|$ denotes the amplitude of the complex scalar a, $\jmath \triangleq \sqrt{-1}$ is the imaginary unit, $\mathbb{E}\{\cdot\}$ is the expectation operator, and $\mathbb{E}_x\{\cdot\}$

specifies the expectation with respect to the random variable x. $\mathbf{x} \sim \mathcal{CN}(\mathbf{a}, \mathbf{A})$ indicates a complex Gaussian random vector with mean $\mathbf{a}$ and covariance matrix $\mathbf{A}$. $O(\cdot)$ and $\Theta(\cdot)$ are the Big-O and the Big-Θ notations, respectively. The rest of the notations used throughout this article and this article's abbreviations are listed in Nomenclature.

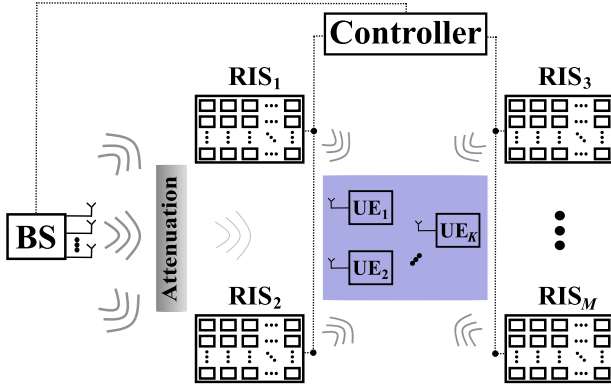## II. CONSIDERED SMART RADIO ENVIRONMENT

In this section, we present the considered smart wireless propagation environment enabled by multiple RISs and introduce an abstract model for the RIS operation and the channel models used throughout this article. We also describe the optimization problem for the joint orchestration of the reflections from the RISs and the BS precoding under investigation.

A. Abstract Modeling of RISs

The typical hardware implementation of RISs is a planar array of ultrathin meta-atoms (also known as unit cells or elements), which have multiple digitalized states corresponding to distinct EM responses [20]. The meta-atoms are usually printed on a dielectric substrate, and their digital control can be achieved by leveraging semiconductor devices, such as positive-intrinsic-negative (PIN) diodes, FETs, and MEMS switches (e.g., varactor diodes) [117], [118].

For quasi-free-space beam manipulation, akin to the main scenario currently considered for RISs in the wireless communications literature [119], [120], fine-grained control over the reflected EM field is essential for accurate beamforming. This fact motivated researchers to rely on meta-atoms of subwavelength size [17], despite inevitable strong mutual coupling between meta-atoms (e.g., when the spacing of adjacent meta-atoms is $d_{\text{RIS}} = \lambda/10$ with λ being the signal wavelength). In contrast, in rich scattering environments, the wave energy is statistically equally spread throughout the wireless medium, and the ensuing ray chaos implies that rays impact the RIS from all possible, rather than in one well-defined, directions. Instead of creating a directive beam, the goal becomes the manipulation of as many ray paths as possible. This manipulation may either aim at tailoring those rays to create constructive interference at a target location or to efficiently stir the field. These manipulations can be efficiently realized with $\lambda/2$-sized meta-atoms, which enable the control of more rays with a fixed amount of electronic components, compared to RISs equipped with their subwavelength counterparts. In addition, mutual coupling among half-wavelength meta-atoms (i.e., $d_{\text{RIS}} = \lambda/2$) is weaker, if not negligible.

In this article, we consider M identical RISs each consisting of $N \triangleq N_h N_v$ meta-atoms placed in groups of N_h in the horizontal dimension, one below the other, such that N_v meta-atoms exist in the vertical dimension. Following recent experimental RIS implementations [121], [122], we adopt the transmission-line circuit model in [123]

Fig. 1. *Considered RIS-empowered smart radio environment comprising an N_T-antenna BS, M identical N-element RISs, and K single-antenna UEs. We assume that the RISs smartly control signal propagation in space in order to boost the performance of the targeted UEs that lie in an area where direct communication from the serving BS is highly attenuated.*

for the meta-atoms. According to this model, each ith element of each mth RIS (denoted as RIS_m from now), with $i = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, M$, behaves like a parallel resonant circuit, and its reflection coefficient $[\phi_m]_i$ determines the fraction of the reflected EM wave due to the impedance discontinuity between the free-space impedance and this element's impedance. Assuming that b is the phase resolution in bits per meta-atom, we consider the following 2^b-element set for each $[\phi_m]_i$:

$$\mathcal{F} \triangleq \left\{ \exp\left(j 2^{1-b} \pi f \right) \right\}_{f=0}^{2^b - 1} \tag{1}$$

which results in a total of 2^{bN} phase profiles (also known as phase configurations) per RIS. It is noted that, in practice, it holds: 1) $|[\phi_m]_i| \leq 1$ and that the amplitude of each reflection coefficient depends both on its phase value and the angle of the impinging wave [20], [123] and 2) $[\phi_m]_i$ is frequency dependent exhibiting a Lorentzian-like frequency response [124] (similar to antennas, an RIS element can be a combination of resonant circuits in series or in parallel [125]).

B. System Model

We consider the downlink wireless communication between a BS equipped with N_T antenna elements and K single-antenna UEs, which is empowered by M identical rectangular RISs. All RISs are assumed to be connected to the same controller, which is responsible for their joint orchestration in conjunction with the multiantenna BS. We consider wired connections between each RIS and the controller, while the latter communicates also with the BS via either a wired or an out-of-band wireless link [21]. As depicted in Fig. 1, we assume that the RISs are placed close to the area where the UEs are located, enabling stronger wireless links (i.e., individual hops) compared

to the case where the UEs are linked to the BS via only direct connections. The geometrical distances of the UE_k-BS, RIS_m-BS, and UE_k-RIS_m links $\forall\, k, m$ are denoted by d_k, d_m, and $d_{k,m}$, respectively, and we, further, assume that $d_{k,m} \ll d_k$. Without loss of generality, we consider the free-space pathloss model according to which $L(d) \triangleq 20 \log_{10}(4\pi d \lambda^{-1})$ in dB, which represents the power loss factor at a certain distance d.

We assume narrowband downlink transmissions where the multiantenna BS deploys the complex-valued $N_T \times K$ precoding matrix $\mathbf{V}$ to multiplex the information symbols for all K UEs, which are included in the vector $\mathbf{q} \triangleq [q_1, q_2, \ldots, q_K]^T$. These symbols, which are herein assumed to be mutually independent, are usually complex-valued and chosen from a discrete modulation set. In a similar manner to [126] and [127], we consider that $[\mathbf{V}]_{:,k} \in \mathcal{V}$ $\forall\, k$, where $\mathcal{V}$ represents the set with the available unit-norm $N_T \times 1$ BS precoding vectors for serving all UEs. By assuming, for simplicity, equal power allocation among the K UE signals,[1] i.e., $\mathbb{E}\{|q_k|^2\} = P/K\ \forall\, k$, the $N_T \times 1$ transmitted signal vector $\mathbf{x} \triangleq \mathbf{Vq}$ is usually constrained as $\mathbb{E}\{\|\mathbf{x}\|^2\} \leq P$ (in our case, the equality holds), where P denotes the total transmit power budget. Using the latter definitions and the following complex-valued N_T-element row vector for the end-to-end channel [14][2]:

$$\mathbf{b}_k \triangleq \left(\sqrt{L(d_k)} \mathbf{h}_k + \sum_{m=1}^{M} \sqrt{L(d_m) L(d_{m,k})} \mathbf{g}_{m,k} \boldsymbol{\Phi}_m \mathbf{H}_m \right) \tag{2}$$

the baseband received signal at each UE_k can be mathematically expressed as follows:

$$\begin{aligned} y_k &\triangleq \mathbf{b}_k \mathbf{x} + n_k \\ &= \mathbf{b}_k [\mathbf{V}]_{:,k} q_k + \sum_{i=1,\, i \neq k}^{K} \mathbf{b}_k [\mathbf{V}]_{:,i} q_i + n_k \end{aligned} \tag{3}$$

where[3] $\mathbf{H}_m \in \mathbb{C}^{N \times N_T}$, $\mathbf{g}_{k,m} \in \mathbb{C}^{1 \times N}$, and $\mathbf{h}_k \in \mathbb{C}^{1 \times N_T}$ represent the channel gain matrices for each of the links RIS_m-BS, UE_k-RIS_m, and UE_k-BS, respectively, while $\boldsymbol{\Phi}_m$

---

[1]Multiuser power control (e.g., [128] and [129]) provides an extra degree of freedom for transmission design optimization and is left for future work.

[2]The cascaded end-to-end channel model of [14] for each BS-RIS_m-UE_k link is almost uniquely deployed in all RIS phase profile design investigations up to date [22]. Very recently (see [130] and references therein), physics-inspired channel models that describe EM wave propagation in various complex media (ranging from free space to rich scattering) are being considered, which cannot be in general decomposed in a cascaded matrices form. Nevertheless, the DRL framework presented in this article can be considered for any RIS-parametrized channel model; its detailed application for the model in [130] is left for future work.

[3]We assume that signals experiencing reflections from more than one RISs are highly attenuated due to the multiplicative pathloss. Therefore, we neglect any inter-RIS channel contribution in the received signal model given by (3).

is defined, using (1), as[4] $\boldsymbol{\Phi}_m \triangleq \operatorname{diag}\{\phi_m\}$ and $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the AWGN. The SINR at each UE_k, which provides a key performance indicator for the considered multiuser wireless communication system, can be easily obtained from (3) as follows:

$$\mathrm{SINR}_k \triangleq \frac{|\mathbf{b}_k[\mathbf{V}]_{:,k}|^2}{\sum_{i=1, \, i \neq k}^{K} |\mathbf{b}_k[\mathbf{V}]_{:,i}|^2 + \frac{K\sigma^2}{P}}. \quad (4)$$

When there is only UE_k present in the RIS-empowered communication system, the SINR_k formula boils down to the expression $\mathrm{SNR}_k \triangleq P|\mathbf{b}[\mathbf{V}]_{:,k}|^2 K^{-1}\sigma^{-2}$, which describes the SNR (in bps/Hz) received at this specific UE.

C. Channel Models

We consider frequency-flat fading channel models for all involved wireless links and assume that channels change independently of one discrete TTI to another.[5] In particular, each channel gain matrix $\mathbf{H}_m$ and each channel gain vector $\mathbf{g}_{m,k}$ are modeled as Ricean faded with factors[6] κ_1 and κ_2, respectively. Each former complex-valued $N \times N_T$ matrix is given $\forall m$ by [138]

$$\mathbf{H}_m \triangleq \sqrt{\frac{\kappa_1}{\kappa_1 + 1}} \mathbf{f}_{\mathrm{RIS}_m}^H \left(\varphi_m^A, \vartheta_m^A\right) \mathbf{f}_{\mathrm{BS}} \left(\varphi_m^D, \vartheta_m^D\right)$$
$$+ \sqrt{\frac{1}{\kappa_1 + 1}} \mathbf{D}_m \quad (5)$$

where φ_m^D and ϑ_m^D denote the azimuth and elevation AoDs of the LOS component leaving the BS, which reaches the RIS_m with the azimuth and elevation AoAs φ_m^A and ϑ_m^A, respectively.[7] The N-element row vector $\mathbf{a}_{\mathrm{RIS}_m}(\varphi, \vartheta)$ for the azimuth and elevation AoAs/AoDs φ and ϑ, respectively, at/from the RIS_m can be expressed, for the considered regular rectangular placement of ideal isotropic meta-atoms at each RIS_m structure, as

$$\mathbf{f}_{\mathrm{RIS}_m}(\varphi, \theta) = \mathbf{f}_{\mathrm{el}}(\theta) \otimes \mathbf{f}_{\mathrm{az}}(\varphi) \quad (6)$$

---

[4]When $d_{\mathrm{RIS}} \ll \lambda/2$ (i.e., nonnegligible mutual coupling), $\boldsymbol{\Phi}_m$ will be a full or banded matrix. For those cases, $\boldsymbol{\Phi}_m$ can be expressed as $\mathbf{C}_m \operatorname{diag}\{\phi_m\}$, where $\mathbf{C}_m$ represents the mutual coupling matrix [131], [132] of the RIS_m structure. This matrix can be then absorbed in the channel matrix $\mathbf{g}_{m,k}$ in (2), rendering the cascaded channel model in this expression mutual-coupling aware. In a similar manner, correlated fading can be included in each involved channel matrix.

[5]To capture the essence of time-varying fading channels, the finite-state Markov channel model can be used [133] that induces channel correlation in time. Such fading conditions can be intuitively exploited to reduce the overhead of RIS optimization, similar to beam tracking schemes (e.g., [134], [135], [136]) in mmWave systems with hybrid analog and digital beamforming. This constitutes an interesting research direction for the DRL-based orchestration of RIS-empowered smart radio environments considered in this article.

[6]Each Ricean factor is defined as the power ratio of the specular component of the specific channel over the random components [137].

[7]All AoAs/AoDs in the array response vectors are taken with respect to the coordinate system defined by the arriving/departing node [136], [139].

where $\mathbf{f}_{\mathrm{el}}(\theta)$ is the complex-valued N_v-dimension elevation steering vector, defined for $\omega \triangleq 2\pi d_{\mathrm{RIS}} \cos(\theta)$ as [140]

$$\mathbf{f}_{\mathrm{el}}(\theta) \triangleq \frac{1}{\sqrt{N_v}} \left[e^{j\omega}, e^{2j\omega}, \dots, e^{N_v j\omega}\right] \quad (7)$$

with d_{RIS} representing the spacing of adjacent meta-atoms in wavelengths in both elevation and azimuth, and $\mathbf{f}_{\mathrm{az}}(\phi)$ is a complex-valued N_h-dimension vector denoting the azimuth steering vector, which is given using the definition $\psi \triangleq 2\pi d_{\mathrm{RIS}} \sin(\phi) \cos(\phi)$ by the expression

$$\mathbf{f}_{\mathrm{az}}(\phi) \triangleq \frac{1}{\sqrt{N_h}} \left[e^{j\psi}, e^{2j\psi}, \dots, e^{N_h j\psi}\right]. \quad (8)$$

The N_T-element row vector $\mathbf{f}_{\mathrm{BS}}(\varphi, \vartheta)$ for the azimuth and elevation AoDs φ and ϑ, respectively, is defined similar to $\mathbf{a}_{\mathrm{RIS}_m}(\varphi, \theta)$ by just replacing d_{RIS} with d_{BS}. Finally, in (5), the $N \times N_T$ matrix $\mathbf{D}_m$ includes the NLOS, i.e., scattering, components of the channel and is modeled as $[\mathbf{D}_m]_{i,j} \sim \mathcal{CN}(0, 1) \ \forall i = 1, 2, \dots, N$ and $\forall j = 1, 2, \dots, N_T$. Each N-element row channel vector $\mathbf{g}_{m,k}$, referring to each UE_k-RIS_m link, is modeled similar to $\mathbf{H}_m$ as follows:

$$\mathbf{g}_{m,k} \triangleq \sqrt{\frac{\kappa_2}{\kappa_2 + 1}} \mathbf{f}_{\mathrm{RIS}_m}(\varphi_{m,k}, \vartheta_{m,k}) + \sqrt{\frac{1}{\kappa_2 + 1}} \mathbf{d}_{m,k} \quad (9)$$

where $\varphi_{m,k}$ and $\vartheta_{m,k}$ denote the azimuth and elevation AoDs of the LOS component between the RIS_m and the position of the UE_k. In addition, the N-element NLOS row vector $\mathbf{d}_{m,k}$ is modeled as $\mathbf{d}_{m,k}^T \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$.

In accordance with the system model presented in Section II-B, we assume that all direct UE_k-BS links are NLOS channels that are power-attenuated by the same attenuation factor $\nu \in [0, 1]$. To this end, each N_T-element row vector $\mathbf{h}_k$ is modeled as Rayleigh faded and is given by

$$\mathbf{h}_k \sim \mathcal{CN}\left(\mathbf{0}_{N_T}, \nu \mathbf{I}_{N_T}\right). \quad (10)$$

Note that the extreme attenuation cases $\nu = 0$ and $\nu = 1$ refer to the absence of the direct links and the nonattenuated direct links, respectively.

D. Orchestration of Smart Radio Environments

The free parameters of the considered RIS-empowered smart radio propagation environment in Fig. 1, i.e., the vectors with the reflection coefficients of the multiple RISs and the BS precoding matrix, are amenable to optimization according to certain performance objectives. For example, such smart environments can be programmed to boost the achievable rate or energy efficiency; increase signal coverage; enable localization, sensing, radio mapping, and secrecy; and provide low EM-field exposure to both intended and unintended UEs in an area(s) of interest [19].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

In this article, we will be using the maximization of the *achievable sum rate* as our exemplary performance metric to devise the core optimization problem to derive RL-based formulations on that (see Section IV-A) and present pertinent results from extensive numerical simulations (see Section VI). Nevertheless, the orchestration of DRL in smart radio environments, as presented in this survey, is readily applicable to any of the aforementioned (or other) network utilities, as long as they can be seen as a function of the responses (i.e., the phases in this article) of the RIS elements. The motivation behind selecting the particular objective is due to its ubiquity in multiuser communications and, subsequently, its extensive use in the comparative literature (see Section IV-B).

The sum-rate performance objective can be mathematically defined using the SINR_k expression in (4) as follows:

$$\mathcal{OP}_1 : \max_{\{\boldsymbol{\phi}_m\}_{m=1}^M, \mathbf{V}} \sum_{k=1}^K \tilde{R}_k \tag{11a}$$

$$\text{s.t. } [\boldsymbol{\phi}_m]_i \in \mathcal{F} \quad \forall m, i, \ [\mathbf{V}]_{:,k} \in \mathcal{V} \ \forall k \tag{11b}$$

$$\tilde{R}_k \geq R_k^{\text{req}} \quad \forall k \tag{11c}$$

where $\tilde{R}_k \triangleq \log_2(1 + \text{SINR}_k)$ represents the instantaneous achievable rate for UE_k's communication link, while R_k^{req} denotes the minimum rate request by this UE. In the sequel, we will treat this design optimization problem both with and without the individual rate constraints in (11c). It noted that, even under the perfect CSI availability case at the controller of the smart radio environment, the latter nonconvex optimization problem with the discrete constraints is hard to solve optimally. It is actually a computationally exhaustive combinatorial problem, whose special cases have been recently suboptimally treated via classical optimization tools in [141] (the case of a single RIS with a multiantenna BS) and via supervised learning in [142] (multiple RISs and a single-antenna BS). In this article, we propose to tackle this problem via the RL methodology, profiting from its inherent capability to offer online learning and optimization in nonstationary conditions (as usually wireless environments are), while avoiding lengthy training collections. In the sequel, we: 1) overview key RL approaches and their application for wireless communications; 2) present our DRL-based formulation for dealing with $\mathcal{OP}_1$; and 3) present the latest advances in DRL schemes for RIS-empowered wireless communication systems.

III. OVERVIEW OF REINFORCEMENT LEARNING

Artificial intelligence and machine learning are very broad domains having already enjoyed great success in wireless communications in general and very recently in representative applications of RIS-empowered communication systems in particular. In this survey paper, the focus is specifically on RL approaches. We invite the readers to consult [143], [144], [145] for tutorials and overviews on deep learning schemes for communication systems based on RISs, which mainly focus on supervised learning methodologies. Herein, we do not attempt to introduce ANNs nor the concept of deep learning; we assume, though, a broad level of familiarity. In fact, we treat ANNs as layered *black box* function approximators that can be trained via the stochastic gradient descent approach. A detailed taxonomy of the RL methods discussed in this section is illustrated in the flowchart of Fig. 2.

The mathematical tool of RL [146] is a machine learning domain that deals with the problem of iteratively maximizing a stochastic objective function that decomposes as a discrete sequence of payoffs. Since such problems are, in general, nonconvex and NP-hard, RL approaches fall under the broad umbrella of nonconvex and stochastic optimization algorithms. The main difference between RL and classical optimization is that, in the former, the problem is assumed to be learnable in the sense that past iterations provide information that can be exploited to provide more efficient actions/performance in the latter stages. The learning part of the RL algorithms is commonly handled by deep, or not, ANNs. In this section, we provide a brief introduction to the RL theory, starting with the main formalism and problem structure, while including useful definitions in formulating RL-based solutions. We also describe how ANNs can be used in the context of DRL to obtain approximate solutions and discuss the main ideas behind the key DRL-based approaches and their evaluation methodology.

A. Markov Decision Processes

In a typical RL problem, a decision-making agent interacts with an environment to achieve a predefined goal. Unlike the paradigms of supervised and unsupervised learning, where datasets must be acquired prior to the training phase, RL constitutes a *closed-loop* approach. An agent continuously observes the information about the state of the system, usually using its measurement collection equipment, and chooses one of the available actions, which is fed back to the system, influencing its evolution in turn. During this interaction, reward signals are communicated by the environment to the agent, which indicates the system's performance at the time of reward collection. The agent is, thus, tasked to learn actions that maximize its cumulative reward signal.

The field of RL has its roots in optimal control, with the concept of the MDP being the cornerstone of the theory behind sequential decision-making [147]. Formally, an MDP is defined as a four-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ with $\mathcal{S}$ and $\mathcal{A}$ being the *state* and *action* spaces, respectively, while $\mathcal{P}$ constitutes the *transition probability* of the environment and $\mathcal{R}(\cdot)$ is the reward function. Note that the action and state spaces can be either continuous or discrete, though each case requires a different treatment, and not
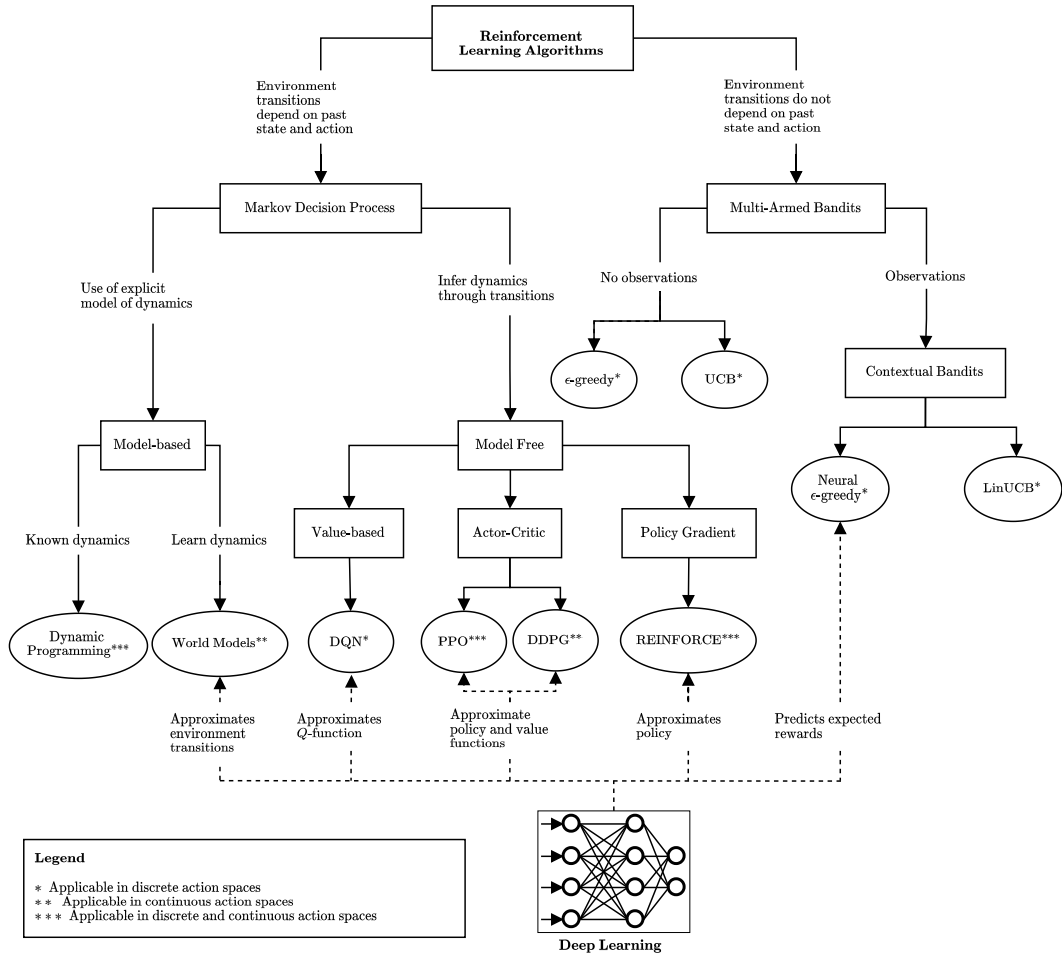
Fig. 2. *Taxonomy of the overviewed RL algorithms in this survey. Ellipses signify RL/DRL algorithms, rectangles denote classes of methods, and text between arrows designates a criterion for the classification. The dashed arrows signify DRL algorithms and explain how deep learning is utilized in each case. The listed algorithms are additionally marked depending on the type of the action space that they admit.*
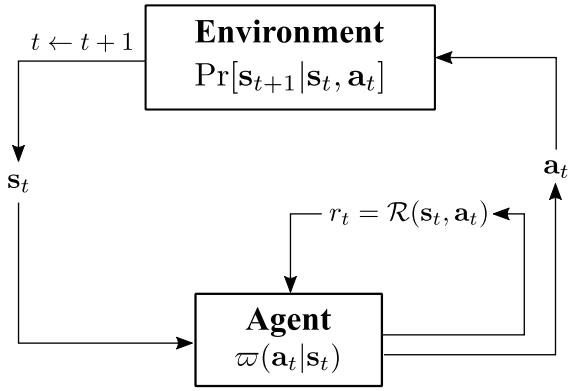
all algorithms are applicable to both. At each (discrete) time step t, the agent: 1) observes the state of the environment $\mathbf{s}_t \in \mathcal{S}$ and 2) selects an action $\mathbf{a}_t \in \mathcal{A}$, with $\mathcal{A}$ including $\mathrm{card}(A)$ discrete actions, which determines the instantaneous reward[8] $r_t \triangleq \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t)$. At the same time, the environment transitions to the next state $\mathbf{s}_{t+1}$ with probability $\mathcal{P} \triangleq \Pr[\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t]$. We will be using the capital letters S_t, A_t, and R_t when referring, respectively, to the random vectors for the state and action, and the random variable for the reward at time t. In addition, without loss of generality, we assume an ordering of the actions within $\mathcal{A}$ using the enumerator operator $\mathbb{I}_\mathcal{A}(\mathbf{a}) \in \{1, 2, \ldots, \mathrm{card}(\mathcal{A})\}$, which returns the index of $\mathbf{a}$ in $\mathcal{A}$ (i.e., it maps an action to its index in the action space). A schematic overview of the MDP formulation is depicted in Fig. 3.

The distinctive characteristic of an MDP is that the following *Markovian property* holds for the transition probability:

$$\Pr[S_{t+1} = \mathbf{s}_{t+1} | \\ S_t = \mathbf{s}_t, A_t = \mathbf{a}_t, S_{t-1} = \mathbf{s}_{t-1}, A_{t-1} = \mathbf{a}_{t-1}, \ldots] \\ = \Pr[S_{t+1} = \mathbf{s}_{t+1} | S_t = \mathbf{s}_t, A_t = \mathbf{a}_t] = \mathcal{P}. \quad (12)$$

This property dictates that the next state at time $t + 1$ depends exclusively on the state-action pair at time t, without being influenced by the history of earlier actions and observations. In effect, the dependence of $\mathbf{s}_{t+1}$ on $\mathbf{s}_t$ describes a time evolution on the environment. The additional dependence on $\mathbf{a}_t$ signifies that the agent is able to influence the transitions, and the Markovian property implies that, by observing $\mathbf{s}_t$, the agent knows all needed to know to predict the next state. In RL implementations, it is convenient to assume the existence of a final time T, which may be infinite. In *finite* MDPs, a subset of the state

---

[8]It is also conceivable for the reward to depend both on the previous state-action pair and the next state. In this case, the reward signal is delayed one time step, i.e., $r_{t+1} = \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$.

Environment
$\Pr[\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t]$

$t \leftarrow t+1$

$\mathbf{s}_t$

$\mathbf{a}_t$

$r_t = \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t)$

Agent
$\varpi(\mathbf{a}_t|\mathbf{s}_t)$

Fig. 3. *Block diagram of an MDP. The agent observes the state of the environment at each time step t and selects an action according to its policy. The action is fed back to environment, which sets the next time step t + 1. A reward is given as feedback to the agent. The reward function can be thought of either as part of the environment or as an external stimulus to the agent.*

space constitutes the *terminal* states, for which no further transitions take place. An agent's sequence of interactions from $t = 0$ to T is termed an *episode*. Typically, multiple episodes are needed for an agent to learn to select good actions.

Formally, the agent selects actions based on a *policy* function $\varpi : \mathcal{S} \to \mathcal{A}$, which describes the mapping from states to actions. This mapping might be nonstationary (i.e., depend on t). Without the loss of generality, policies can be assumed to be stochastic. In this case, $\varpi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ describes the probability of selecting an action given a state. We will use the notation $\varpi(\mathbf{a}|\mathbf{s})$ when referring to likelihood values of actions and ϖ when referring to a policy as a function. Hence, the objective of typical RL problem formulations can be expressed as finding the optimal policy that maximizes the expected *return* from some time step t

$$\mathcal{OP}_2 : \; \varpi^* \triangleq \underset{\varpi}{\arg\max} \, \mathbb{E}_{\varpi} \{G_t^{\varpi}\} \qquad (13)$$

where the return G_t^{ϖ} is defined as the discounted sum of rewards, given that the agent takes actions following policy ϖ:

$$G_t^{\varpi} \triangleq \sum_{k=0}^{T-t} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}^{\varpi}. \qquad (14)$$

In the above expression, $\gamma \in (0, 1]$ denotes the *discount factor*, which makes the sum of rewards finite even in the infinite case. Conceptually, the discount factor imposes a preference on larger immediate rewards versus future ones. Typical discount values reside in the interval $[0.9, 0.99]$, aiming to promote long-term gains instead of myopic strategies.

The typical MDP formalism assumes that the agent is capable to observe the state of the environment instantly. This is evidently an impractical assumption for the majority of real-world applications, and it is especially problematic for wireless communication systems. The theory of POMDPs, apart from the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, further supposes an observation space and a probability density function of collecting a particular observation given the true state. The agent sees only a distorted observation of the true state, making the problem of learning the value functions exceedingly hard. Even though there are exact and approximate methods devoted specifically to POMDPs, in practice, algorithms that have been proposed in the context of MDPs are also applied to this category, with the assumption that the noisy observations contain enough useful information to sufficiently estimate the target functions. DRL approaches, which deploy ANNs to learn useful representations of the input states and will be described in the sequel, tackle the problem of partial observability. They usually equip the agents with memory units (e.g., by using recurrent networks) to aid them to distinguish the true state of the environments they reside in.

B. Value Functions

It is common for RL algorithms to keep a notion of value functions that estimate the benefit of an agent observing a state or an action-state pair. Specifically, for a given policy ϖ, the *value function* at any state $\mathbf{s} \in \mathcal{S}$ is mathematically defined as

$$\begin{aligned} V^{\varpi}(\mathbf{s}) &\triangleq \mathbb{E}_{\varpi} \{G_t^{\varpi}|S_t = \mathbf{s}\} \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \varpi(\mathbf{a}|\mathbf{s}) \sum_{\mathbf{s'} \in \mathcal{S}} \Pr[\mathbf{s'}|\mathbf{s}, \mathbf{a}] \left(\mathcal{R}(\mathbf{s}, \mathbf{a}, \mathbf{s'}) + \gamma V^{\varpi}(\mathbf{s'})\right) \end{aligned}$$
$$(15)$$

which describes the expected return when the environment at time step t is at state $\mathbf{s}$ and the agent follows policy ϖ. In the latter expression, discrete action and state spaces were assumed, but the notation can be extended to continuous spaces. Accordingly, the *action value function* at any state $\mathbf{s} \in \mathcal{S}$ and action $\mathbf{a} \in \mathcal{A}$ is given by the expression

$$\begin{aligned} Q^{\varpi}(\mathbf{s}, \mathbf{a}) &\triangleq \mathbb{E}_{\varpi} \{G_t^{\varpi}|S_t = \mathbf{s}, A_t = \mathbf{a}\} \\ &= \sum_{\mathbf{s'} \in \mathcal{S}} \Pr[\mathbf{s'}|\mathbf{s}, \mathbf{a}] \\ &\quad \times \left(\mathcal{R}(\mathbf{s}, \mathbf{a}, \mathbf{s'}) + \gamma \sum_{\mathbf{a'} \in \mathcal{A}} \varpi(\mathbf{a'}|\mathbf{s}) Q^{\varpi}(\mathbf{s'}, \mathbf{a'})\right) \end{aligned}$$
$$(16)$$

which concerns the expected value when following policy ϖ and selecting action $\mathbf{a}$ when observing state $\mathbf{s}$. The latter two value functions are connected by the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

following relationship:

$$V^{\varpi}(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} \varpi(\mathbf{a}|\mathbf{s}) Q^{\varpi}(\mathbf{s}, \mathbf{a}). \qquad (17)$$

In order to solve $\mathcal{OP}_2$ appearing in (13), it suffices to obtain the optimal value functions, i.e., find the maximum values for each state (or action-state pair) under any policy. By exploiting the recursive structure of (15) and (16), the optimal value functions for any state and any state-action pair can be derived through Bellman's *principle of optimality*, which is expressed via the Bellman equations

$$
\begin{aligned}
V^*(\mathbf{s}) = \max_{\mathbf{a}} \sum_{\mathbf{a} \in \mathcal{A}} \varpi(\mathbf{a}|\mathbf{s}) \\
\times \sum_{\mathbf{s'} \in \mathcal{S}} \Pr[\mathbf{s'}|\mathbf{s}, \mathbf{a}] \left(\mathcal{R}(\mathbf{s}, \mathbf{a}, \mathbf{s'}) + \gamma V^*(\mathbf{s'}) \right)
\end{aligned}
\qquad (18)
$$

and

$$Q^*(\mathbf{s}, \mathbf{a}) = \sum_{\mathbf{s'} \in \mathcal{S}} \Pr[\mathbf{s'}|\mathbf{s}, \mathbf{a}] \left(\mathcal{R}(\mathbf{s}, \mathbf{a}, \mathbf{s'}) + \gamma \max_{\mathbf{a'} \in \mathcal{A}} Q^*(\mathbf{s'}, \mathbf{a'}) \right). \qquad (19)$$

These equations define a pair of dynamic programming problems that can be solved by backward induction. Upon attaining the optimal value functions, it is straightforward to define the optimal policy as the one that selects the action, which maximizes the expected return via the Q-function, i.e.,

$$\mathcal{OP}_3: \ \varpi^*(\mathbf{a}|\mathbf{s}) \triangleq \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}} Q^*(\mathbf{s}, \mathbf{a}). \qquad (20)$$

This optimal solution of the MDP suffers, however, from two important drawbacks. First, the transition probabilities of the environment are assumed to be known to the decision-maker (i.e., *model-based RL*), which is restrictive for real-world applications. Second, the dynamic programming solution requires iterating over all possible states or state-action pairs, which is intractable for most demanding practical problems.

C. Deep Reinforcement Learning Algorithms

Recent advances in the RL field aim to approximate different functions of the MDP formalism in order to derive inexact, but efficient, solutions. Even though some approaches may target learning directly the transition function of the environment (e.g., "World Models" of [148]), the vast majority of the RL methods employ a *model-free* strategy, performing estimates using the experience collected by the agent. In general, a DRL algorithm is primarily equipped with a policy function that may be directly parameterized (i.e., using an ANN) or defined indirectly (i.e., through the maximization of value functions that are parameterized instead).

Algorithm 1 Training Procedure of a DRL Agent

Require: MDP description $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, final time step T, and for the agent: initial ANN parameters $\mathbf{w}$, policy $\varpi_{\mathbf{w}}$, collection policy $\hat{\varpi}_{\mathbf{w}}$, objective function $J(\cdot)$, and update interval t'.
1: Observe initial state $\mathbf{s}_1$ from the environment.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:     Use $\hat{\varpi}_{\mathbf{w}}$ to determine action $\mathbf{a}_t$.
4:     Feed $\mathbf{a}_t$ to the environment to observe $\mathbf{s}_{t+1}$ and r_t.
5:     Store experience $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t)$ to a set $\mathcal{D}$.
6:     **if** $\mathrm{mod}(t, t') = 0$ **then**
7:         Use $\mathcal{D}$ to compute $\mathbf{w}^*$ optimizing $J(\mathbf{w})$.
8:         Update as $\varpi_{\mathbf{w}} \leftarrow \varpi_{\mathbf{w}^*}$ and $\hat{\varpi}_{\mathbf{w}} \leftarrow \hat{\varpi}_{\mathbf{w}^*}$.
9:     **end if**
10: **end for**
11: **return** Learned policy $\varpi_{\mathbf{w}}$.

During training, the agent uses a *collection policy* to interact with the MDP, as well as amasses and stores tuples of observations, actions, and rewards. At certain intervals (potentially at every time step t), a variant of gradient descent/ascent algorithm is invoked to learn the parameters of the underlying ANNs so that a pertinent objective function over the collected data is optimized. The collection policy may coincide with the policy intended to be learned (*on-policy* algorithms) or may deviate from it to encourage better exploration (*off-policy* algorithms). A generalized description of the DRL training process of an agent is provided in Algorithm 1. The algorithm is purposely given in an abstract form since the exact scheme varies according to the specifics of each method, but, in principle, all algorithms incorporate ANNs designed to observe state vectors and output "quality assessments" over actions.

Ordinarily, DRL approaches are categorized either as value-based, policy-gradients, or a combination of the two (actor–critic methods). We next present the basic components of certain important methods that are commonly used in the field of wireless communications and RIS-empowered communication systems. A detailed walk-through of the state of the art in DRL approaches can be found in the recent survey [149].

1) Value-Based Approaches: The DQN algorithm [150] constitutes the most well-established value-based DRL method for discrete action spaces. An ANN is employed to estimate the optimal Q-function of the underlying environment using data collected from the agent's past interactions. This Q network receives as input a state representation and outputs predictions of the Q-values for each action (i.e., the output layer has $\mathrm{card}(\mathcal{A})$ units). To estimate the action value function, the agent collects experience tuples of the form $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$, which are stored in a replay buffer. At every iteration, experience tuples are sampled in minibatches $\mathcal{D}$ from the buffer, and the network updates its current estimate of the Q-value for

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

the pair $(\mathbf{s}_t, \mathbf{a}_t)$ based on the TD error; this error indicates the deviation of the predicted optimal Q-value for the next state from the estimated Q-value of the current action-state pair. By denoting the ANN as $Q_{\mathbf{w}}$, with $\mathbf{w}$ being the concatenated parameter vector with the real-valued network's weights, the loss function to be optimized is given by

$$
\mathcal{L}(\mathbf{w}) \triangleq \sum_{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t) \in \mathcal{D}} (r_t + \gamma \max_{\mathbf{a} \in \mathcal{A}} Q_{\mathbf{w}}(\mathbf{s}_{t+1}, \mathbf{a}) \\
- Q_{\mathbf{w}}(\mathbf{s}_t, \mathbf{a}_t))^2 \tag{21}
$$

where the summation is over the experience in the minibatch. The training of the ANN can be handled by any variation of the stochastic gradient descent (e.g., Adam [151] and RMSprop [152]), following the generic gradient update rule for $k = 0, 1, \ldots$.

$$
\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \alpha \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_k) \tag{22}
$$

where $\alpha \in (0, 1]$ is the *learning rate* hyperparameter. This training procedure is known to converge, under mild assumptions, to the optimal Q-function up to a statistical error that reflects the fundamental difficulty of the problem [153].

DQN is an *off-policy* algorithm according to which the policy the agent uses for collecting experience is not the same with the one to be ultimately learned. The learned policy is effectively the selection of the action that maximizes the predicted Q-values, i.e., as in (20). During training, the ϵ-*greedy* heuristic is used, in which a random action is occasionally selected instead of the one dictated by the maximum Q-value

$$
\mathbf{a}_t \leftarrow \begin{cases} \mathcal{U}(\mathcal{A}) & \text{with probability } \epsilon \\ \text{argmax}_{a \in \mathcal{A}} Q(\mathbf{s}_t, a) & \text{with probability } 1 - \epsilon. \end{cases} \tag{23}
$$

In this expression, $\mathcal{U}(\mathcal{A})$ represents the selection operation from the discrete set $\mathcal{A}$ with probability $1/\text{card}(\mathcal{A})$, and the value of ϵ is used to balance the *exploration-exploitation* dilemma: By only relying on already explored strategies, the agent is prone to suboptimal performance. Thus, a degree of random exploration is required. On the contrary, by disregarding previously gained information, the algorithm will take more ineffective actions than those necessary for training. Typically, ϵ takes values in [0.01, 0.2], while it is also common to start the training procedure with a larger value and then decrease it iteratively.

In [150], the efficacy of DQN was demonstrated by training it to achieve human-like performance in Atari games. An important insight from that work is that DQN training can be fairly unstable due to frequent updates of the network's parameters. Effectively, the same network is generating the next state target Q-values that are used

in updating its current Q-values, which is known to be prone to oscillations or divergence. To account on that, an identical network to $Q_{\mathbf{w}}$, dubbed the target network $Q_{\mathbf{w}^-}$ with separate parameters $\mathbf{w}^-$, was employed to predict the current estimate of the optimal Q-value [i.e., (21)'s term $r_t + \gamma \max_{\mathbf{a}} Q_{\mathbf{w}}(\mathbf{s}_{t+1}, \mathbf{a})$]. While the original Q network is updated during every time step t, the target network is meant to be updated at a lower frequency, and its new weights are calculated as a "soft-copy" from the current value of $\mathbf{w}$ according to a "temperature" hyperparameter $\hat{\tau}$

$$
\mathbf{w}_{k+1}^- \leftarrow (1 - \hat{\tau})\mathbf{w}_k^- + \hat{\tau}\mathbf{w}_k. \tag{24}
$$

This trick offers increased stability when the estimates over the states change slowly enough for the policy to be able to adapt to the current estimations. Often, the accompanying practice of clipping the network's gradient values within a specified range is employed to impose more restraint changes on the network's predictions.

2) Policy-Gradient Approaches: The PG-based approaches aim to directly find a policy that maximizes the expected sum of rewards in episodic tasks. In practice, they adopt a parametric approximation of the policy function, i.e., an ANN that maps states to actions. For notation purposes, let us rewrite $\mathcal{OP}_2$'s objective in (13) as follows:

$$
J_{\text{PG}}(\mathbf{w}) \triangleq \mathbb{E}_{\tau \sim \varpi_{\mathbf{w}}} \left\{ \sum_{t=1}^T R_t \right\} = \mathbb{E}_{\tau \sim \varpi_{\mathbf{w}}} \left\{ G_t^{\varpi_{\mathbf{w}}} \right\} \tag{25}
$$

where the dependence on the network's parameter vector $\mathbf{w}$ is made explicit. The term τ denotes a *trajectory*, i.e., a sequence of state-action pairs throughout the episode, generated by following $\varpi_{\mathbf{w}}$. The optimization of $J(\mathbf{w})$ is carried out through the gradient ascent iteration $(k = 1, 2, \ldots)$

$$
\mathbf{w}_k \leftarrow \mathbf{w}_k + \alpha \nabla_{\mathbf{w}} J_{\text{PG}}(\mathbf{w}_k). \tag{26}
$$

The PG theorem [154] states that the stochastic gradient of the objective function can be computed as

$$
\nabla_{\mathbf{w}} J_{\text{PG}}(\mathbf{w}) = \mathbb{E}_{\tau \sim \varpi_{\mathbf{w}}} \left\{ \sum_{t=1}^T \nabla_{\mathbf{w}} G_t^{\varpi_{\mathbf{w}}} \log \left(\varpi_{\mathbf{w}}(\mathbf{a}_t | \mathbf{s}_t) \right) \right\}. \tag{27}
$$

To approximate this expectation, Monte Carlo sampling can be used resulting in the *on-policy* algorithm REINFORCE [155]. Note that the theorem does not consider the discounted case of infinite MDPs (since the training takes place after the end of the episode), but PG algorithms may use discounts to compute the returns of finite MDPs in practice.

3) Actor–Critic Approaches: A problem with Monte Carlo estimates is that they have high variance with respect to the true gradient of the objective function. One common technique to reduce the resulting noisy gradient updates is to adopt an *advantage* function in place of $G_t^{\varpi \mathbf{w}}$ in (27). The most prominent example is the *advantage value* function, which quantifies the benefit of selecting action $\mathbf{a}_t$ at state $\mathbf{s}_t$ when compared to the (average) value of the state. Specifically, this function is defined as

$$\mathsf{A}(\mathbf{s}_t, \mathbf{a}_t) \triangleq Q^{\varpi \mathbf{w}}(\mathbf{s}_t, \mathbf{a}_t) - V^{\varpi \mathbf{w}}(\mathbf{s}_t)$$
$$= r_{t+1} + \gamma V^{\varpi \mathbf{w}}(\mathbf{s}_{t+1}) - V^{\varpi \mathbf{w}}(\mathbf{s}_t). \qquad (28)$$

Again, the true value function is unknown, but it can be estimated from experience with an ANN. Hence, the actor–critic framework employs two distinct networks: the *actor* that learns the optimal policy using a PG-like method and the *critic* that is trained to predict the value of the state, allowing for (28) to be substituted in (27).

The PPO algorithm [156] is a DRL method that further deals with the problem of noisy updates by imposing a proximity constraint. According to this algorithm, the gradient values are clipped within a trust region, which discourages the likelihood of selecting any specific action under the new policy. The goal of this strategy is not to deviate substantially from the action's likelihood given by the old policy. Apart from being a relatively straightforward method to implement, PPO has also shown great performance in multiagent training scenarios [157].

Designed specifically to handle continuous action spaces, the DDPG algorithm [158] is another established actor–critic method. It uses a network (critic) to estimate Q-values similar to DQN. Dealing with the problem of finding the action that maximizes the value is computationally expensive in the continuous domain. Therefore, a separate action-selection network (actor) is used to determine the value of the selected action. The latter networks are trained concurrently in an off-policy manner to minimize their respective objectives, each one having its parameters fixed when computing the gradients of its counterpart. Target networks are employed in both subproblems to stabilize the training process, as proposed by the DQN algorithm.

Remark: The value-based methods are, in general, more straightforward to implement and arguably more interpretable since they provide evaluations of the different states (and actions) of the MDP. On the contrary, the need for computing one value per state-action pair makes them inefficient in (combinatorially) large and discrete action spaces, even if discretization is applied. On the other hand, PGs can adapt to both of those cases since a parametric probability distribution is adopted that can allow for efficient action sampling. Despite that, PG updates are usually less sample efficient since trajectories of arbitrary length must be collected through Monte Carlo sampling, and each gradient update affects only the network's belief on the action selected. This is in contrast to value-based methods that may update the estimation of a state's

value for all actions simultaneously. Another difference between these two categories arises from the on-/off-policy distinction of their algorithms. PGs, being on-policy, are trained strictly when exploring the environment. Contrarily, value-based methods are flexible to also be trained completely off-policy (i.e., on previously collected MDP transitions), which might be crucial when online querying of the environment is expensive or unavailable. Finally, actor–critic methods are designed to capitalize on the benefits of the above two categories since they employ both a value and a policy network. In fact, current state-of-the-art DRL methods belong to this class of algorithms, notwithstanding their increased computational needs and added complexity. A schematic overview of the three main DRL approaches discussed in this survey is given in Fig. 4. The diagrams are kept simple by excluding less important aspects of the presented methods for the sake of clarity.
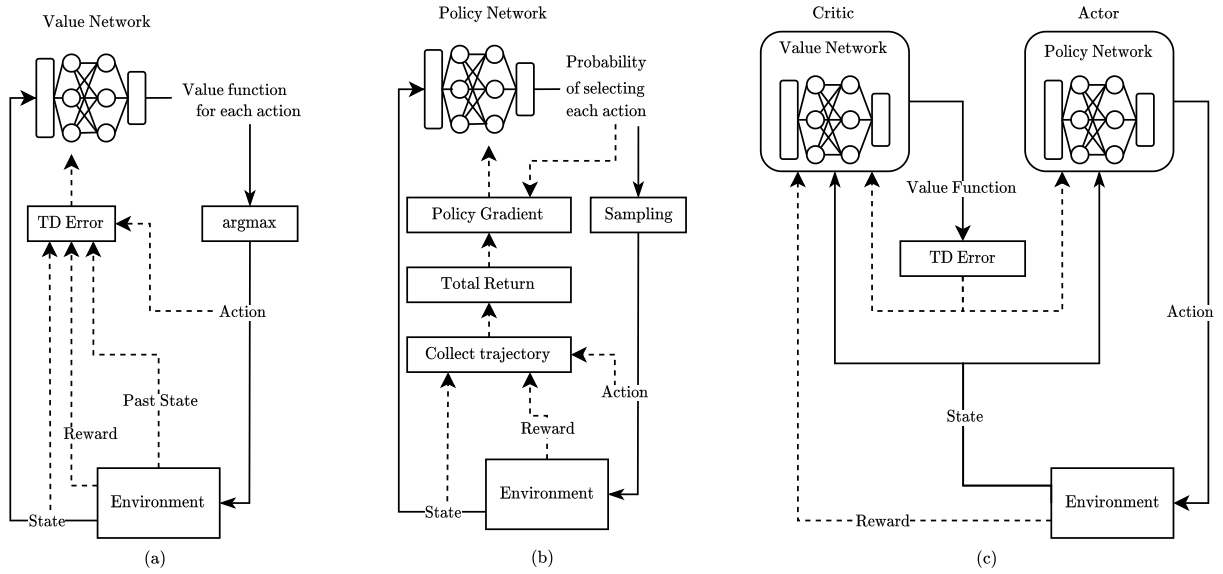
D. Multiarmed Bandits

As previously mentioned, one of the basic principles of the MDP formalism is that the action selected by an agent influences the next state of the environment. Hereinafter, we consider the case where the agent still selects an action to maximize a cumulative reward signal and potentially receives observations, but its actions have no impact on the evolution of the environment. Such sequential decision-making problems are commonly referred to as *multiarmed bandits*. Concretely, at each time step, a decision-maker needs to select an action from a discrete action set. The environment internally keeps a separate probability distribution of the payoff (reward) for each action and provides the agent with a reward sampled from the distribution corresponding to the selected action. The agent is again tasked with maximizing the expected (potentially discounted) reward within a single episode. We note that, in the current version of formalism, there are no states present. As a result, the notion of policy is not applicable. Instead, the agent needs to find the most rewarding action(s), as quickly as possible, to maximize its expected payoff.

The multiarmed bandits' problem formulation exemplifies the exploration-exploitation dilemma of MDP-based approaches since the agent needs to try different actions to infer the optimal one(s). Most commonly, the algorithms applicable to this case first keep track of a running average of the achieved reward for each action. This running average serves as an estimate of its true expected value and is expressed as

$$\mathsf{G}_{t+1}(\mathbf{a}) \triangleq \frac{1}{\mathsf{N}_t(\mathbf{a})} \sum_{i=1}^{\mathsf{N}_t(\mathbf{a})} r_i^{\mathbf{a}} = \mathsf{G}_t(\mathbf{a}) + \frac{1}{\mathsf{N}_t(\mathbf{a})} \left[r_{\mathsf{N}_t(\mathbf{a})} - \mathsf{G}_t(\mathbf{a}) \right] \qquad (29)$$

where $\mathsf{N}_t(\mathbf{a})$ is the number of times action $\mathbf{a}$ has been selected up to time t. Note that the MDP formalism has been slightly altered by defining the per-action return and by denoting with $r_i^{\mathbf{a}}$ the reward associated with $\mathbf{a}$ when the action is selected for the ith time. This learned information

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

Fig. 4. *Illustration of the three main approaches of MDP-based DRL methods. Solid lines denote input/output variables of the ANNs and the flow of information during the MDP cycle. Dashed lines denote the flow of information during the training process in each approach. (a) Value-based learning. (b) PG. (c) Actor–critic.*

is further combined with an exploration strategy. We briefly describe two of the most common methods.

1) The ϵ-**greedy** strategy, which was previously described. We explicitly redefine the strategy to incorporate the current multiarmed bandits' framework as

$$\mathbf{a}_t \leftarrow \begin{cases} \mathcal{U}(\mathcal{A}) & \text{with probability } \epsilon \\ \text{argmax}_{\mathbf{a} \in \mathcal{A}} \, \mathsf{G}_t(\mathbf{a}) & \text{with probability } 1 - \epsilon. \end{cases} \tag{30}$$

2) The **UCB** heuristic, which takes into account both the maximum estimated values and the uncertainty in those estimates

$$\mathbf{a}_t \leftarrow \text{argmax}_{\mathbf{a} \in \mathcal{A}} \left\{ \mathsf{G}_t(\mathbf{a}) + c\sqrt{\frac{\ln(t)}{\mathsf{N}_t(\mathbf{a})}} \right\} \tag{31}$$

where $c > 0$ controls the width of the confidence interval and, hence, the amount of exploration. For completeness, when $\mathsf{N}_t(a) = 0$, only the confidence bound term is set to 0. It is usually expected from UCB to perform better than the naive ϵ-greedy approach since the former has the advantage of an uncertainty-guided exploration, instead of the random strategy of the latter.

It is noted that the multiarmed bandits' problem concerns both the cases of stationary and nonstationary distributions over the actions. In the latter case, the agent would benefit by giving preference to more recent observations. As a result, the running average of (29) can be modified to an exponential recency-weighted average [146, Ch. 2].

A more general formulation of the bandits' problem admits observations by the agent. In this *contextual bandits* case, the agent receives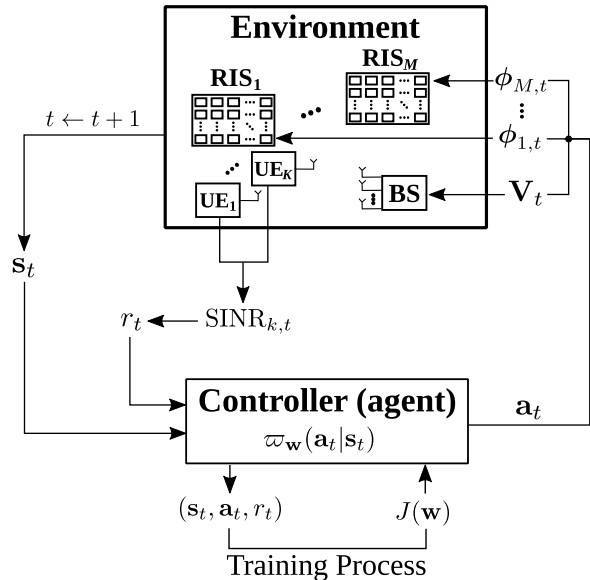 an observation vector (termed the context) from the environment prior to deciding each action, with the intent of being guided to the optimal selection strategy. Therefore, the notion of policies applies to contextual bandits, but, since there are no transition probabilities between states and actions, contextual bandits' problems are considered relatively easier than the full-MDP ones. It is also worth mentioning that the observation vector is not assumed to fully represent the internal state of the system, but it is rather expected to have a relationship with the reward values. Conceptually, such bandit settings can be thought of as episodic MDP problems with a trajectory of length 1 (i.e., $T = 1$). Due to that, standard DRL algorithms can be applied in theory, but bandit-tailored methods are more suited since the former algorithmic approaches are unnecessarily complex.

Indicative in this contextual bandits category is the LinUCB algorithm [159], which assumes a linear dependence between observations and actions. Specifically, the expected reward for each action, which depends on the current observation (i.e., context), is modeled as a linear relationship between an unknown coefficient vector and the observation vector. That coefficient vector is estimated through ridge regression on past observation-rewards data, and an upper bound of the expected payoff is derived. As such, the UCB action-selection strategy is finally employed to determine the chosen action. Unsurprisingly, the assumption of linearity between observations and rewards can be restrictive in many challenging problems; this means that the potential representations of the coefficient vectors are limited. To remedy for that, Riquelme *et al.* [160] proposed the neural linear bandits' algorithm, which utilizes an encoding neural network to learn the coefficient vector instead of ridge regression. The network was trained as a regressor to predict the reward given an observation. To attain the encoding vector, the outputs

of the final hidden layer were extracted, i.e., the original output layer was dropped. A Thomson sampling action-selection strategy was employed instead of UCB, but the approach can be trivially imported to the LinUCB setting.

Another straightforward approach to incorporate deep learning into the contextual bandits' formulation would be via a *reward-prediction* ANN. Such a network accepts an observation vector as input at time t and outputs as many real numbers as the number of possible actions for the agent. Those numbers are the expected reward values for each action, i.e., $G_t(\mathbf{a})$ $\forall \mathbf{a} \in \mathcal{A}$. Thus, the ϵ-greedy strategy of (30) can be utilized during the learning process. To this end, the network can be trained using tuples of observation and reward, and the mse between predicted and true rewards can be used as the objective function. Note that the mse is computed only with respect to the output neuron that corresponds to the selected action and not to the whole output vector of the network. We will be referring to this approach as the "neural ϵ-greedy" algorithm in the sequel.

Although MDP and POMDP formulations are the most popular for wireless communications up to date [103], [161], [162], [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], the multiarmed bandits' formulations have a substantial proposition of value, despite their simplicity. The multiarmed bandits' algorithms are more intuitive and explainable compared to DRL methods. Their converge properties are rather easier to be studied theoretically [174], and their effectiveness can be demonstrated empirically [175] while being able to utilize the immense representational power of deep learning [176]. In the section with the numerical results, different multiarmed bandits' algorithms for the orchestration of the considered RIS-empowered smart wireless environments will be evaluated.

E. Algorithmic Evaluation

The theoretical analysis of DRL algorithms is an active topic of research. Convergence results have been established for well-known methods under fairly mild conditions (e.g., [153] for DQN, [154] for PG, and [177] for actor–critic), but they may be asymptotic; consider only specific families of MDPs (e.g., [177]) or impose constraints on the underlying functions. The fundamental obstacle in attaining general results is that the MDP formalism makes almost no assumptions about the underlying problem. Therefore, the rate and point of convergence are dependent on the characteristics/difficulty of the application.

A most useful evaluation practice is that of experimental evaluation. DRL algorithms are commonly applied to benchmark environments of different levels of difficulty with various characteristics. Different metrics can be used to assess their performance, namely, the cumulative reward achieved and the number of samples (expressed in time steps) required up to a certain performance threshold; the latter provides a notion of *sample efficiency*. In general, off-policy algorithms are expected to be more sample-efficient

Fig. 5. *Block diagram of the proposed DRL formulation for the considered RIS-empowered smart radio environment in Section II-B with the sum-rate maximization design objective.*

since they are able to iterate over previous experiences multiple times.

In terms of execution time per iteration, the DRL algorithms are commonly treated as constant time. In general, the time complexity of a forward pass in an ANN(s), which is required for a prediction, is of the order of the number of the training parameters in $\mathbf{w}$ with an added $\Theta(\mathrm{card}(\mathcal{A}))$ cost due to the argmax operation on discrete spaces or an equivalent cost of the sampling from a distribution in the continuous case. The training involves taking a gradient step and backpropagating the gradient through the network, which entails the same complexity as the forward pass, assuming $O(1)$ batch sizes of potential minibatches.

IV. DRL-ORCHESTRATED SMART RADIO ENVIRONMENTS

In this section, we first present a DRL-based formulation of the design objective introduced in Section II for RIS-empowered smart wireless environments and discuss its applicability for various other performance objectives. Second, we provide a detailed overview of the available DRL approaches for wireless communication systems, including RISs.

A. DRL Formulation

Following the theoretical foundations of (D)RL, as presented in Section III, for the considered system model in Section II-B and its achievable sum-rate maximization design problem given by (11), we map the parameters of the wireless communication system with the DRL ingredients, as illustrated in Fig. 5 and discussed in the sequel.

1) The controller, which is assigned the role to orchestrate the multiantenna BS, the multiple RISs, and the multiple UEs as per (11)'s objective, is the *agent*

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

in the DRL formulation, being specifically capable of deciding at each time step t (e.g., per channel coherence time or performance-based requests) the phase profiles for all RISs and the precoding matrix of the BS. We redefine the former as $\{\phi_{m,t}\}_{m=1}^{M}$ and the latter as $\mathbf{V}_t$ to explicitly indicate their dependence on the instant t.

2) The possible *actions* that the controller (i.e., the agent) can take are determined by the feasible combinations of the free design parameters of the wireless system, which constitutes the *action space* $\mathcal{A}$. For the considered system, this space includes $[\phi_m]_i \in \mathcal{F}$ $\forall m, i$ and $[\mathbf{V}]_{:,k} \in \mathcal{V}$ $\forall k$, resulting in $\text{card}(\mathcal{A}) = 2^{bN_{\text{tot}}} K \text{card}(\mathcal{V})$. In addition, an action at any time t can be defined by the following complex-valued[9] $KN_{\text{tot}}N_T$-element column vector

$$\mathbf{a}_t \triangleq [\phi_{1,t}, \phi_{2,t}, \ldots, \phi_{M,t}, \text{vec}(\mathbf{V}_t)]^T. \quad (32)$$

3) The wireless propagation environment corresponds to the *environment*,[10] whose state can be naturally mapped to the channel gain coefficients that change per channel coherence time. Consequently, the controller (i.e., the agent) can observe the environment via performing/collecting estimates of the elements of the channel matrices. The state of the environment at a time instant t can be defined by the following complex-valued column vector with dimensionality $\dim(\mathbf{s}_t) \triangleq N_{\text{tot}}(K + N_T) + KN_T$:

$$\mathbf{s}_t \triangleq [\text{vec}(\tilde{\mathbf{H}}_t); \mathbf{g}_{m,k,t}^T; \mathbf{h}_{k,t}^T]^T \quad (33)$$

where we have used the definitions

$$\tilde{\mathbf{H}}_t \triangleq [\mathbf{H}_{1,t}, \mathbf{H}_{2,t}, \ldots, \mathbf{H}_{M,t}] \in \mathbb{C}^{N_{\text{tot}} \times N_T} \quad (34)$$
$$\tilde{\mathbf{g}}_t \triangleq [\mathbf{g}_{1,1,t}^T, \mathbf{g}_{1,2,t}^T, \ldots, \mathbf{g}_{M,K,t}^T] \in \mathbb{C}^{KN_{\text{tot}} \times 1} \quad (35)$$
$$\tilde{\mathbf{h}}_t \triangleq [\mathbf{h}_{1,t}^T, \mathbf{h}_{2,t}^T, \ldots, \mathbf{h}_{K,t}^T] \in \mathbb{C}^{KN_T \times 1}. \quad (36)$$

In practice, the controller possesses the estimate $\hat{\mathbf{s}}_t$ of $\mathbf{s}_t$.

4) The controller is capable of inferring the outcome of its actions on the wireless environment at any time instant t. This is usually quantified via the *reward* for this instant, which constitutes the feedback of the controller's actions from the environment that can be collected by the controller. For the considered sum-rate maximization objective in $\mathcal{OP}_1$, the agent (i.e., the controller) is assumed to be capable of

collecting the SINR measurements of each UE_k at any instant t, denoted by $\text{SINR}_{k,t}$. Focusing first on the case where the individual UE rate constraints in (11c) are excluded from $\mathcal{OP}_1$, the agent uses the latter measurements to calculate the following reward:

$$r_t \triangleq \sum_{k=1}^{K} \tilde{R}_{k,t}. \quad (37)$$

Note that, since the controller gathers the CSI state estimate $\hat{\mathbf{s}}_t$ of the environment at any instant t, the SINR values and, hence, the reward for this instant can be estimated without the intervention of the UEs (i.e., without the need to collect feedback from them). Nevertheless, relying on a UE reporting mechanism to generate the reward signal at the agent ensures that the pragmatic reward is captured, which, in the previous case, would be prone to imperfections due to the collection of partial or erroneous CSI (or even insufficient modeling of the channel and the control operation).

For the case where $\mathcal{OP}_1$ is considered exactly as it appears in (11) (i.e., including the (11c) constraints), the reward function needs to be revised to penalize actions that result in states where any of the individual rate requirements is not satisfied. There exist various techniques for achieving this goal, which actually depends on various aspects of the MDP. In this article, we choose to reshape the reward function to return -1 for each UE_k whose individual SINR_k did not satisfy the specified request (i.e., when the specific constraint in (11c) was not met). Following this formulation, the agent was considered to compute the reward

$$r_t = \begin{cases} -\sum_{k=1}^{K} \mathbb{1}_{(\tilde{R}_k < R_k^{\text{req}})}, & \text{if } \exists k : \tilde{R}_k < R_k^{\text{req}} \\ \sum_{k=1}^{K} \tilde{R}_k, & \text{otherwise} \end{cases} \quad (38)$$

where $\mathbb{1}_{(\cdot)}$ denotes the indicator function. It is noted that, while maximizing the reward in (38) indeed maximizes $\mathcal{OP}_1$, the achieved average rewards can no longer be interpreted as achievable sum rates. This happens because, in the definition of (38), negative scores affect the resulted metric. Nevertheless, positive values for rewards indicate that the individual UE rate requests are successfully met; in this case, the resulting reward equals the optimized achievable sum-rate performance.

5) The RIS phase profiles and the BS precoding matrix chosen by the controller at any time instant t (i.e., the agent's action) will have no effect on the channel gain matrices (i.e., the state of the environment) of the next time instant $t + 1$. Together with our assumption for i.i.d. channel realizations (in the time dimension) in Section II-B, it yields for our DRL formulation

---

[9]The implementations of the RL algorithms are not designed to handle complex-valued numbers. Due to this, it is common to concatenate the real and imaginary parts (or magnitude and phase) of complex-valued vectors into real vectors, having effectively a double dimension.

[10]The environment in the context of RL includes any aspect of a system that is not under the direct control of an agent. For example, in our system model, UE properties, such as locations, mobility, rate requests, or other constraints, may be considered parts of the environment.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

that the *Markovian transition probability* in (12) is independent of the previous state $\mathbf{s}_t$ and action $\mathbf{a}_t$, i.e., it holds: $\Pr[\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t] = \Pr[\mathbf{s}_{t+1}]$. This constitutes a special case of the MDP formulation that better adheres to the multiarmed bandits' paradigm, as discussed in Section III-D. It is noted, however, that, when time-correlated channels are involved (see, e.g., the model in [133]), the channel state at each time instance is dependent on the preceding channel state(s). Considering that the channel state at each time $t + 1$ depends only on the channel state at t, as an indicative example of time-correlated channels, the Markovian transition probability is expressed as $\Pr[\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t] = \Pr[\mathbf{s}_{t+1}|\mathbf{s}_t]$. Even though MDP-based algorithms are inherently capable of capturing such time dependencies, the multiarmed bandits' outlook can still be adopted. In particular, any algorithm falling into the latter category needs to be applied in ways such that it updates the expected reward estimation by emphasizing the most recent observations. In Section VI-F, the learning capability of the proposed controllers based on multiarmed bandits will be numerically investigated over i.i.d. and time-correlated channels.

From a system architecture and operation perspective, training a DRL-based agent entails a nonnegligible infrastructure and computational overhead. At every time instant, CSI measurements need to be performed and collected by the controller. Leveraging the uplink/downlink channel reciprocity in time-division duplexing systems, channel acquisition can be implemented by transmitting pilot signals from the UEs in the uplink and applying estimation techniques for the direct and the RIS-assisted channels at the BS and/or the controller sides [118]. With this information at the controller's disposal, the sum-rate maximizing phase profiles for the RISs and the BS precoding matrix are computed and shared with the involved devices. Finally, the UEs are required to measure their individual SINR values and feedback them to the controller, facilitating its training process in a dynamic manner. The complete DRL training procedure for a time window of length T is described in Algorithm 2. Note that, due to the consideration of i.i.d. state transitions, $\hat{\mathbf{s}}_{t+1}$ needs not to be stored in any experience tuple (see Step 7), as in Algorithm 1, since it does not provide any instructive information in learning the environment's dynamics or determining an appropriate action. Moreover, the practical aspect of obtaining a CSI estimate $\hat{\mathbf{s}}_t$ in place of $\mathbf{s}_t$ poses the current formulation as an POMDP problem, rather than the simple MDP case. In practice, however, ANN-based algorithms are capable of attaining notable performance, even in the presence of observational noise (i.e., noisy CSI).

After the completion of the training process described in Algorithm 2, the controller configures the elements of the smart wireless environment for downlink communication according to the learned policy $\varpi_\mathbf{w}$. To this end, all

---

Algorithm 2 DRL-Based Training Solving $\mathcal{OP}_1$ in (11)

Require: Number of UEs K, number of RISs M, initial ANN parameters $\mathbf{w}$, policy $\varpi_\mathbf{w}$, collection policy $\hat{\varpi}_\mathbf{w}$, objective function $J(\cdot)$, final time step T, and update interval t'.

1: **for** $t = 1, 2, \ldots, T$ **do**
2:     Estimate $\mathbf{h}_{k,t} \ \forall k = 1, 2, \ldots, K$.
3:     Estimate $\mathbf{H}_{m,t}$ and $\mathbf{g}_{m,k,t} \ \forall m = 1, 2, \ldots, M$ and $\forall k = 1, 2, \ldots, K$.
4:     The controller formulates the state vector $\hat{\mathbf{s}}_t$ similar to (33) using the channel estimates in Steps 2 and 3.
5:     The controller decides the action $\mathbf{a}_t$ (i.e., $\phi_{m,t}$ $\forall m$ and $\mathbf{V}_t$) using its policy $\hat{\varpi}_\mathbf{w}(\mathbf{a}_t|\hat{\mathbf{s}}_t)$ and shares these settings with the RISs and BS.
6:     Each UE_k measures $\mathrm{SINR}_{k,t}$ and sends it to the controller.
7:     Using Step 6, the controller computes the reward r_t and stores the experience tuple $(\hat{\mathbf{s}}_t, \mathbf{a}_t, r_t)$ to a set $\mathcal{D}$.
8:     **if** $\mathrm{mod}(t, t') = 0$ **then**
9:         The controller uses $\mathcal{D}$ to compute $\mathbf{w}^*$ that optimizes $J(\mathbf{w})$.
10:        It then performs the updates $\varpi_\mathbf{w} \leftarrow \varpi_{\mathbf{w}^*}$ and $\hat{\varpi}_\mathbf{w} \leftarrow \hat{\varpi}_{\mathbf{w}^*}$.
11:    **end if**
12: **end for**
13: **return** Learned policy $\varpi_\mathbf{w}$.

---

steps of this algorithm, except Steps 6–10 referring to the ANN training, can be used for the orchestration with the difference being that the controller uses $\varpi_\mathbf{w}$ instead of $\hat{\varpi}_\mathbf{w}$ in Steps 5 as its policy. Note that the SINR collection is no longer needed at the ANN deployment phase since no training takes place.

Although our DRL formulation considers the sum rate as the performance objective, it is general enough to be equivalently applied to any desired performance metric [19] that is a direct function of the current channel coefficients and RIS phase profiles (e.g., energy efficiency, low EM-field exposure, secrecy rate, NMSE, and localization/sensing accuracy). It is also important to mention that (D)RL formulations are by no means limited to observing CSI and controlling standalone RISs. A wide variety of more elaborate problems can be conceptualized in which the agent observes different kinds of information (e.g., current UE positions) and controls diverse elements of the environment (e.g., power allocation and UAVs with mounted RISs). In the following, we discuss the state of the art in DRL approaches for RIS-empowered wireless communication systems.

B. Literature Overview

In this section, we provide a brief overview of the literature on DRL-based orchestration for RIS-empowered smart

---

radio environments. In the largest portion of the overview, the presented approaches consider a system model similar to the one considered in this article and detailed in Section II. More specifically, most of the literature employs a DRL agent that is trained using CSI observations and is assigned the role to learn to configure the phase shifts of the RIS (possibly along with other system parameters), thus making the underlying MDP formulations adhere to the general formalism described in Section IV-A. Each of the discussed studies considers a different variation that is based on the peculiarities of the proposed system model. We begin by describing the approaches that are trained to optimize the UE(s)' achievable rates and are, thus, more closely related to this article's design formulation.

A DRL formulation similar to that of Section IV-A was adopted in [103]. This work considered continuous RIS phase profiles and precoding matrices, and as a result, the DDPG algorithm was deployed to learn the optimal configurations. The transmission power at every time step was embedded in the observation vector, and the underlying policy network was extended with an additional layer that ensured the satisfaction of the problem's power constraint. In [167], the same design formulation with the extension of power allocation was transferred in a multi-RIS communication system with mobile UEs operating at the THz frequency band. It was shown that DDPG is capable to find the RIS phase configurations, as well as the BS precoding vectors and power allocation strategies that can compensate for the severe attenuation of THz signal propagation, outperforming an alternate beamforming benchmark in terms of total throughput. Gao *et al.* [161] considered a more elaborate system model that consisted of the consecutive subproblems of estimating the UEs' mobility patterns, finding their optimal allocation into clusters, and computing the RIS phase configuration. Different machine learning algorithms were devised for each case, with DQN set as the RIS orchestrating approach. The overall algorithm received the results of the aforementioned tasks as inputs, apart from the CSI-based observation vector defined in Section IV-A, in its aim to maximize the sum-rate performance.

Some works in the area addressed the impracticality of assuming full CSI knowledge when deploying DRL-based RIS controllers; this issue will be revisited in the open challenges in Section V-B. The overhead of the channel estimation process in RIS-empowered communication systems was the main focus of the DRL approach in [162]. A hybrid RIS was adopted with a small percentage of active elements able to perform channel sensing. As a result, the observation space was comprised of estimates of the true channel state, making the problem a POMDP. The authors trained the ubiquitous DQN agent (which is also considered in the numerical evaluation in Section VI) with quantized rewards ($+1$ or -1 depending on whether the current (single) UE rate exceeded the previous value), demonstrating that DRL methods are applicable even in partial channel knowledge, while requiring minimal training overhead.

In a similar manner, Feng *et al.* [166] investigated the DDPG agent in a scenario with a mobile UE without relying on CSI knowledge. Instead, the action and reward of the past time step constituted the observation vector. The DRL algorithm was responsible for finding the (continuous) RIS phase shifts that maximize the SNR, while the precoder was derived through MRT [178]. A multicellular multi-RIS system model was investigated in [170], where DRL methods were trained to decide on the BS transmit power and precoder, as well as the RIS phase configuration that maximizes the UEs' SINR values. The RIS controller (which was a DQN controller) was assumed to have access to only local channel states, and the proposed DRL schemes were evaluated under imperfect CSI conditions. In Section VI, we follow an approach similar to [166] in order to evaluate DRL methodologies with and without CSI-based observations by devising a strong LOS scenario with mobile UEs (i.e., time-evolved channels).

In the context of this survey, we have focused on the sum-rate maximization as the objective of the (D)RL formulations. However, a number of works in the RIS field dealt with the EE or the secrecy rate design problems [163], [165], [168], [169]. As discussed in Section V-A, the general formulation in this work can be modified to account for other objectives through the pertinent definition of the reward function. In [168], RIS elements that the controller can turn on or off, as part of the action space of the RL formulation, were further considered. In [165], a multi-UAV system model, in which the UAV facilitates the downlink transmission in clusters of UEs alongside an RIS, was investigated. The employed PPO algorithm observed channel states and controlled the power allocation at the UAVs. The DDPG and PPO algorithms were compared against each other, both in centralized and decentralized (i.e., MAML) system architectures. The EE design problem was also studied in [169] for MISO NOMA networks with mobile UEs. Apart from the power allocation, the considered design problem accounted for displacements of the RIS, and both were embedded into the action space of the RL formulation. The proposed approach leveraged the D^3QN variant of the DQN algorithm, in which the Q network was tasked to learn both the value function and the advantage function to compose the Q-value for each state-action pair, paired with a decaying ϵ-greedy exploration policy. Finally, Yang *et al.* [163] adopted the DQN algorithm under the previously described design formulation, with the purpose of maximizing the secrecy rate of the legitimate UEs under an eavesdropping scenario.

Although the DRL-based design formulation in this work is quite general to account for a large number of applications, it is worth mentioning that more elaborate systems do not adhere to this framework. Like [165], the works [171], [172], [173] considered mobile RISs and opted for traditional optimization algorithms for their orchestration, which differs from the general formulation presented herein. Specifically, Al-Hilo *et al.* [171] dealt

Table 1 Taxonomy of the State of the Art of DRL-Based Approaches for RIS-Empowered Smart Radio Environments

Reference	Method	Action	State	Reward (Objective)
[161]	DQN	RIS phase profile, power allocation	Past action	Difference between current and previous instantaneous rate
[162]	DQN	RIS phase profile	Estimated CSI	Rate (quantized)
[103]	DDPG	RIS phase profile, beamforming vector	CSI, past action, transmit power	Sum rate
[163]	DQN (extended)	RIS phase profile, beamforming vector	Past CSI, past rates	Secrecy rate, QoS constraints
[164]	PG-Based	Single meta-atom phase response	RIS phase profile, index of meta-atom, time counter	Sensing error
[165]	DDPG, PPO	RIS phase profiles, power allocation	Reflected channels	Energy efficiency
[166]	DDPG	RIS phase profile	Past SNR value, past action	SNR
[167]	DDPG	Phase profiles of multiple RISs, beamforming matrix	CSI, past action	Sum rate, power constraints
[168]	DQN	RIS phase profile, ON/OFF meta-atoms, transmit power allocation	Beamforming vectors of UEs, RIS energy level	Energy efficiency
[169]	D^{3}QN	RIS phase profile, transmit power allocation, RIS placement	Positions of UEs, past phase profile, past power allocation, past RIS position	Energy efficiency
[170]	DQN	Phase profiles of multiple RISs, transmit power allocation, beamforming vector	Local and neighboring channel powers, neighboring sum rates	Sum rate minus interference
[171]	PPO	Selection of vehicles to be served	Vehicle positions, velocities, rates	Minimum average rate
[172]	PPO	RIS phase profile, UAV height	UAV altitude, rate, and AoI of IoT devices	Negative sum AoI
[173]	Quantile Regression	UAV location	Received signal power	Total received data during transmission

with the problem of scheduling mobile UEs. In [173], the RIS was positioned onto a UAV that was proposed to be controlled by the quantile regression DRL algorithm [179], targeting coverage improvement and data-rate boosting. The altitude of the UAV along with the RIS phase configuration to optimize the AoI across IoT devices was considered in [172]. Diversely, Hu *et al.* [164] proposed a DRL-based sensing system in an RIS-empowered wireless environment, for which the MDP formulation involved sequentially selecting the phase shifts of the RIS to identify the intended objects. A PG-based approach was designed, which worked in conjunction with a supervised learning model that was used for object detection.

It is worth mentioning that the choice between the most popular DRL-based agents that are utilized by the wireless communications community (i.e., DQN and DDPG) depends foremost on the nature of the RIS elements used in the underlying system models since the aforementioned algorithms are designed specifically for either discrete or continuous action spaces, respectively. A more detailed taxonomy of the proposed DRL methods is presented in Table 1 in terms of their respective RL formulations. The majority of the literature is focused on MISO systems in the presence of a single RIS and multiple UEs, whereas many works incorporate additional tasks to the action space of the DRL formulation (e.g., BS spatial processing and power allocation). Finally, we would like to note that many of the considered system models in the aforementioned previous art are characterized by i.i.d. channel realizations (i.e., states of the environment) and available actions (namely, RIS phase configuration and BS precoding matrix selection), which does not influence future states of the system. Therefore, they adhere to the multiarmed bandits formalism firstly considered in this article although they have been examined under the MDP-based framework. In Section VI that follows including the numerical

evaluations, we make use of the newly introduced formulation to assess the performance of a contextual bandits DRL agent, which will be shown to perform similar to the DQN algorithm in a number of scenarios, and with lower implementation complexity.

V. PRACTICAL CONSIDERATIONS AND OPEN CHALLENGES

Without intending to be exhaustive within the fast-evolving area of DRL-based orchestration for RIS-empowered smart wireless environments, we present, in this section, some key practical considerations and open challenges arising from our design formulation previously described in Section IV-A.

A. Extensions With Practical Considerations

In various practical cases, time-varying channels are more likely to appear instead of the considered, in Section IV-A, as well as in the majority of the available literature, i.i.d. channel realizations. Time evolution reinstates the dependence of each state s_{t+1} on s_t (and potentially on more previous states) in the transition probability, implying that the underlying DRL-based algorithms must be able to store and be trained on transitions (i.e., tuples consisting of actions, observations, and rewards) of a greater length than 1, as appearing in Algorithm 2.

In another respect, in the considered orchestration formulation, a centralized approach has been adopted, in which the controller is the sole agent, with the intention to facilitate all UEs. It is, however, conceivable for such a setup to be extended and studied in more realistic distributed manners, e.g.: 1) with the UEs being the agents, having potentially conflicting goals and 2) with multiple controllers, thus again multiple agents, each controlling a single RIS or a subset of the RISs or groups of meta-atoms

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

in different RISs. The latter considerations will naturally lead to MAML and game-theoretic formulations, which are out of the scope of this article, although they constitute exciting practical considerations. In the latter case, the activation and assignment of RISs to controllers may be dynamic, yielding self-organizing schemes of spatially distributed RISs for meeting demanding design objectives, such as interference management/mitigation.

B. Open Research Challenges

Apart from the lack of theoretical guarantees on the performance of the presented (D)RL approaches in Section III, there are plentiful of other open problems and challenges related to their application. Works described in the relevant literature overview in Section IV-B address some of the following issues; however, we list the ones we consider most important here to provide a synopsis of the potential research directions.

1) Environmental Observations and Control Overhead: The RL formulations assume that the agent is capable of querying the environment about observations and reward signals. In wireless communication systems, this entails, on the one hand, the availability of sensing equipment and the adoption of relevant signal processing schemes in order to primarily measure the channel states and the desired performance indicators (e.g., communication rates). On the other hand, separate channels to and from the system controller need to be established for the control information exchange. Evidently, there are tradeoffs between the costs associated with the quality and frequency of CSI estimation, as well as the achievable performance with the specific RL implementation. Another aspect to consider is the channel occupancy due to the potential control and CSI observation exchange processes (in-band or out-of-band [19], [21]) and the associated degradation in achievable performance. Up to date, in the vast majority of the DRL-based orchestration approaches for RIS-empowered smart wireless environments, it is implied that penalties induced by such system control operations are negligible. While this is a reasonable assumption for certain studies, the quantification of such overheads on the overall performance is, to our view, a critical issue and presents an important research direction, especially in complex system setups, such as the multi-RIS setup considered in this article.

2) Simulations Versus Real-World Deployment: From a practical standpoint, the numerical evaluation of wireless communication schemes is easier to be performed via simulations of the signal propagation environment. It is, therefore, a long-standing question of whether the employed channel models and operation setups capture all important aspects of the underlying physical systems. This is especially true for the emerging field of RISs and RIS-empowered smart wireless environments, for which the exact EM models are well less understood at the moment (see [130] and references therein). As a result, the transfer of policies learned in simulated environments to physical systems needs to be treated with care by the RIS community. Crucially, some important progress has been made in the DRL area. For example, the domain adaptation problem has been recently dealt with for a robotic arm manipulation in [180], where further randomness was imposed on the simulated environments.

3) Combinatorially Large Action Spaces: The RISs to be deployed in future wireless communication systems are expected to contain a large number of phase-tunable unit elements. The same is true about the number of the BS antenna elements and, thus, the dimension of the precoding vectors. Recall, for our DRL formulation in Section IV-A that, when the controller of multiple RISs plays the role of the agent, the cardinality of the action space is exponential to the total number of the RIS meta-atoms. It is clear that the problem of attaining optimal performance may become intractable for large enough RISs (or BS antenna elements) or may require considerable training periods. To remedy for that, the deployed orchestration algorithms may need to resort to domain-specific optimization parts along with DRL components in order to reduce the exponential search space. Another algorithmic approach would be to consider DRL algorithms that are specifically tailored to large action spaces (e.g., [181] and [182]) or actions consisting of binary vectors [183] since most practical RIS hardware design involve low phase resolution meta-atoms (e.g., $b = 1$ and 2 [20]). One solution, when 1-bit phase resolution meta-atoms are considered, is to model the RIS phase configuration as a binary-element vector, which would result in an action space that is linear to the total number of RIS elements. Leveraging this modeling that enables individual tuning of each meta-atom, modified versions of DQN and DDPG can be devised.

4) RIS Operation Capabilities and Computational Requirements: An important issue to consider in the envisioned RIS-empowered smart wireless environments is the physical location of the controller for the RISs and, consequently, the part(s) of the network connected intelligence equipped with computational capabilities. Taking into consideration the demanding requirements of ANN training (i.e., GPU), it is rather assumed that the required computation may be offloaded to a dedicated mobile edge computing server. This is, admittedly, the most general approach though it requires establishing pertinent communication links, as discussed in Section IV-A. Conversely, combining the DRL-based intelligent controller of the RIS with the actual RIS may greatly reduce the overhead of the control information exchange by also taking advantage of the recently proposed hybrid RIS hardware architectures with embedded sensing capabilities [96], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111]. In such configurations of smart wireless environments, far fewer control information exchanges for realizing DRL will be required, albeit with the added cost of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

the computationally and sensingly autonomous RIS. Again, such overheads are not yet extensively studied, especially for multi-RIS deployment scenarios.

VI. NUMERICAL RESULTS AND DISCUSSION

Having conducted computer simulations for the multi-RIS-enabled smart wireless environment included in Section II, we present, in this section, the performance evaluation of the proposed DRL methodology, presented in detail in Section IV-A, which has as an orchestration objective the sum-rate maximization criterion expressed in (11). We have elaborated on a number of practical aspects for the design of the proposed (D)RL algorithms and the benchmark schemes, including hyperparameter selection and performance evaluation strategies.

We also present a *novel* multiarmed bandits methodology to solve the design problem at hand, leveraging the assumption of i.i.d. transition probabilities discussed in Section IV-A (due to the considered channel model, which assumes independence of the channel realizations in time). Recall that, since the RIS phase profiles and the BS precoding matrix, which need to be designed at every coherent time step by the orchestration controller, affect only the current time frame, the online orchestration problem conceptually decomposes to individual sum-rate maximization problems per time step. This has the effect of making the optimal policy "myopic," i.e., the greedy strategy that selects the RIS phase profiles and the BS precoder that maximize the sum rate in each current channel realization vector $\mathbf{s}_t$. Consequently, the design of the optimization problem can be reduced to predicting the sum rate/reward, given the current CSI state $\mathbf{s}_t$, for all available actions. This outlook naturally casts the orchestration problem of the multi-RIS smart wireless environment to a contextual bandits' setting, in which the goal is to obtain a model capable to capture the distribution of the reward signals. To solve this problem, we employ the *neural ϵ-greedy* agent, which was briefly discussed in Section III-D and presented in detail in the following.

A. Proposed and Benchmark Methods

1) Neural ϵ-Greedy: The core of the neural ϵ-greedy algorithm is a reward-prediction ANN, represented by $\hat{G}_{\mathbf{w}}(\cdot)$ and parameterized by the network's weights $\mathbf{w}$, which is tasked to predict the expected sum-rate performance values for all RIS phase profiles and BS precoding matrices, given a CSI observation. In particular, at each coherent time instant t, the CSI observation vector $\mathbf{s}_t \in \mathcal{S}$ is fed to the ANN, whose output $\hat{\mathbf{r}}_t \triangleq \hat{G}_{\mathbf{w}}(\mathbf{s}_t)$ is a real-valued vector with $\mathrm{card}(\mathcal{A})$ elements. The ith element, with $i = 1, 2, \ldots, \mathrm{card}(\mathcal{A})$, of the latter vector signifies the estimation of the expected reward if the ith action of the action space $\mathcal{A}$ is selected. The ϵ-greedy exploration strategy selects the action that corresponds to the maximum element of $\hat{\mathbf{r}}_t$ with probability $1 - \epsilon$.
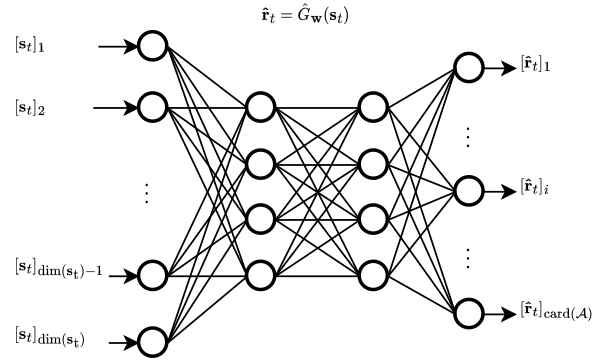
Fig. 6. *Reward-prediction ANN used by the neural ϵ-greedy algorithm. The network receives an observation $\mathbf{s}_t$ as input and outputs a vector of reward predictions $\hat{\mathbf{r}}_t$, the ith element of which corresponds to an estimation of the expected reward if the action with index i is selected.*

The proposed reward-prediction ANN for the orchestration controller (i.e., the agent in our DRL formulation) is illustrated in Fig. 6, and its exact architecture is provided in Table 2. As shown, the ANN is a feedforward neural network with two convolutional layers (each preceding a max-pooling operation), followed by two fully connected layers with rectified linear unit (ReLU) activation functions and a final output layer. The dropout technique was used for regularization after each layer apart from the last. The convolutional structure of the ANN is motivated by the fact that the dimensionality $\dim(\mathbf{s}_t)$ of the state vector is linear to the size and the number of RISs, leading to high-dimensional input vectors. Furthermore, since adjacent elements in the channel vectors/matrices correspond to physical properties of adjacent antenna elements, spatial correlation [178] is expected to be present in the ANN's input state vector, which is appropriately treated using convolution and max-pooling operations. The values for the hyperparameters of the proposed neural ϵ-greedy bandits' algorithm are given in Table 2. Their values have been set empirically, upon exploring a small number of different combinations of them. Notice that the value of the probability ϵ of selecting a random action is uncharacteristically high. This is in light of the very stochastic nature of the wireless channel realizations with the purpose of making the controller better explore the action space and discouraging it to converge early to potentially suboptimal actions.

Table 2 Hyperparameter Values for the Neural ϵ-Greedy Algorithm

ϵ	0.3
t'	32
α	0.001
Convolutional layer units	64, 64
Convolution kernel width	5, 5
Hidden fully connected layer units	32, 32
Max-pooling patch width	4, 4
Dropout probability	0.2

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

Algorithm 3 Neural ϵ-Greedy Algorithm Solving $\mathcal{OP}_1$ in (11)

Require: Probability ϵ of selecting a random action, reward-prediction ANN $\hat{G}_\mathbf{w}$ with initial parameters $\mathbf{w}$, learning rate α, final time step T, and update interval t'.
1: Initialize $\mathcal{D} \leftarrow \emptyset$.
2: Observe initial state $\mathbf{s}_1$ from the environment.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Get the prediction $\hat{\mathbf{r}}_t = \hat{G}_\mathbf{w}(\mathbf{s}_t)$.
5: $\mathbf{a}_t \leftarrow \begin{cases} \mathcal{U}(\mathcal{A}) & \text{with probability } \epsilon \\ \arg\max_{\mathbf{a} \in \mathcal{A}}\{[\hat{\mathbf{r}}_t]_{\mathbb{I}_\mathcal{A}(\mathbf{a})}\} & \text{with probability } 1 - \epsilon \end{cases}$
6: Feed $\mathbf{a}_t$ to the environment to observe $\mathbf{s}_{t+1}$ and r_t.
7: Store $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r_t)\}$.
8: **if** $\mod(t, t') = 0$ **then**
9:    Compute $\hat{\mathcal{L}}(\mathbf{w})$ from (39).
10:   Perform $\mathbf{w}' \leftarrow \mathbf{w} - \alpha\nabla_\mathbf{w}\hat{\mathcal{L}}(\mathbf{w})$.
11:   Update $\hat{G}_\mathbf{w} \leftarrow \hat{G}_{\mathbf{w}'}$.
12:   Reset $\mathcal{D} \leftarrow \emptyset$.
13: **end if**
14: **end for**
15: **return** Trained network $\hat{G}_\mathbf{w}$.

Table 3 Hyperparameter Values for the DQN Algorithm

ϵ	0.3
t'	1
α	0.002
Batch size	128
Gradient clipping threshold	$(-1000, 1000)$
Target network update interval	100
Target network temperature $\hat{\tau}$	0.18
Convolutional layer units	64, 64
Convolution kernel width	5, 5
Hidden fully connected layer units	32, 32
Max-pooling patch width	4, 4
Dropout probability	0.2

The training of the controller involves fitting the ANN on the most recent batch of collected experiences. Assuming a batch size of t' coherent time steps, the controller spends them to explore the environment and store experience tuples of the form $(\mathbf{s}, \mathbf{a}, r)$ into a batch set $\mathcal{D}$. At the t'th time step, the ANN is partially trained via single-step gradient descent to minimize the mse between the actual and the predicted rewards on the batch. Note, however, that, although the network outputs a vector of predictions, only the true (scalar) reward (i.e., the one corresponding to the action ultimately selected) is made known to the controller. As a result, the residual differences of the mse calculation are taken only with respect to the output neuron that corresponds to the selected action. Concretely, the loss function of the ANN was designed as

$$\hat{\mathcal{L}}(\mathbf{w}) \triangleq \sum_{(\mathbf{s},\mathbf{a},r) \in \mathcal{D}} \left(\left[\hat{G}_\mathbf{w}(\mathbf{s})\right]_{\mathbb{I}_\mathcal{A}(\mathbf{a})} - r \right)^2 \qquad (39)$$

where $\hat{\mathbf{r}} = \hat{G}_\mathbf{w}(\mathbf{s})$ is the ANN's prediction vector with the $\text{card}(\mathcal{A})$ reward values; therefore, $[\hat{\mathbf{r}}]_{\mathbb{I}_\mathcal{A}(\mathbf{a})}$ is a scalar prediction of the reward for the action $\mathbf{s}$ actually taken. The proposed algorithmic steps for training the ANN-based reward-prediction controller are included in Algorithm 3.

2) DQN: The performance of the proposed multiarmed bandits' algorithm has been compared to the well-known DQN algorithm, which was described in Section III-C1, and serves as a representative MDP-based DRL approach. Note that, even though the formulation of Section IV-A lends itself to a contextual bandits application scenario, MDP-based approaches are still applicable since bandits' problems can be seen as special cases of MDPs with time horizon $T = 1$. The intention behind such a comparison is to numerically investigate whether the added

computational complexity, in terms of ANN requirements and hyperparameter tuning, brought by a state-of-the-art MDP-based DRL algorithm offers any benefits over the more simplistic approach based on multiarmed bandits.

For a fair complexity comparison with the proposed DRL approach, we incorporated in DQN a Q network identical to the reward-prediction ANN of the neural ϵ-greedy algorithm. It is noted here that DQN uses two independent copies of this network. The considered parameters for the DQN algorithm are given in Table 3. The values of the parameters for both methods were determined by a hyperoptimization tuning algorithm over a set of possible values (specifically, Bayesian optimization), considering a simple scenario for computational purposes.

3) Random and Optimal Schemes: To ascertain the applicability of the proposed and benchmark DRL methods as intelligent orchestration controllers for the investigated RIS-empowered smart radio environments, we make use of the following baseline and upper bound approaches: 1) the "Random policy" that chooses the RIS phase profiles and the BS precoders at random at every coherence time interval and 2) the "optimal policy" that computes, at every time step, the best action from the available ones, through exhaustively searching among all available RIS phase profiles and BS precoding matrices for the ones that maximize the achievable sum-rate performance.

4) UCB: Acquiring observations in the form of CSI estimations may be a demanding task, in terms of time and computational resources, in certain practical wireless communication systems, especially during the deployment phase of any (D)RL algorithm (in contrast to the training phase). For this reason, it is important to investigate whether the considered (D)RL-based approaches are capable of retaining a remarkable portion of their performance in application scenarios without the demanding availability of CSI observations; for example, using only the feedback of SINR measurements from the UEs for the reward computation. The proposed multiarmed bandits' formulation is actually applicable for this limited-feedback case (i.e., K real values instead of $N_{\text{tot}}(K + N_T) + K N_T$ complex channel coefficients), as described in Section III-D. In fact, the SINR values directly yield the reward via (37) or (38), which implies that, for this simple RL formulation, only the

Table 4 Parameters' Setting Used in the Simulation Results

BS coordinates (m)	$(10, 5, 2)$
RIS_1 coordinates (m)	$(7.5, 13, 2)$
RIS_2 coordinates (m)	$(12.5, 13, 2)$
UE_1 coordinates (m)	$(8.775, 14.394, 1.634)$
UE_2 coordinates (m)	$(9.648, 13.281, 1.632)$
N_{T}	4
K	2
$\text{card}(\mathcal{V})$	4
κ_1, κ_2	30 dB
P	40 dBm
σ_k^2 (equal for all UEs)	-110 dBm
f	5 GHz
N_{group}	16
N_{tot}	$\{32, 48, 64, 80, 160\}$
N_{control}	$\{2, 4, 6, 8, 10\}$
$\text{card}(\mathcal{A})$	$\{16, 64, 256, 1024, 4096\}$
$\dim(\mathbf{s}_t)$	$\{400, 784, 1168, 1552, 1936\}$

individual rewards (i.e., rates and, consequently, SINRs) determine the decision-making process (i.e., the selected actions). Note that, for the proposed neural ϵ-greedy algorithm, we have used the contextual version of multiarmed bandits, which requires both the CSI observations and the feedback of the individual rewards for learning the actions' selection policy. Similarly, it holds for the DQN algorithm.

For the purpose of comparing the proposed neural ϵ-greedy algorithm against an observation-free approach, we employ the simple UCB heuristic strategy given by (31), which selects the action that has the highest upper bound estimate of the average gain associated with it. The running averages for each action are then computed via (29) and are initialized randomly to make the algorithm capable to explore the environment for the first few coherent time steps. The UCB algorithm requires only a single parameter c that controls the width of the confidence intervals, which was empirically set to 0.6 for our experimentation process. It is noted that this parameter depends, in general, on various features of the problem (e.g., scale and variances of the reward values).

B. Simulation Setup

In all figures that follow, we have considered a BS with $N_T = 4$ antennas, whose downlink communication with $K = 2$ single-antenna UEs was enabled by $M = 2$ identical RISs, which were placed close to the UEs. The positions of the latter were sampled at random and kept fixed throughout the numerical evaluation process. To lay emphasis on the capability of RISs to enable wireless communications, we considered the presence of a blocker that obstructed each BS-UE$_k$ link, i.e., the respective coefficients of the direct channel vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ were set to zero. The rest of the parameters used in the performance evaluations are included in Table 4.

As discussed in Section II, the design objective of the proposed multi-RIS-empowered smart wireless environment is to maximize the sum-rate performance, i.e., solve $\mathcal{OP}_1$ in (11), by allowing the orchestration controller

(i.e., the agent) to observe the channel matrices per coherent time interval and, accordingly, configure the profiles of the RISs along with the selection of the BS precoding matrix. Cases with and without the constraint in (11c) for the individual rate requests have been investigated. We have considered a number of setups varying the total number of RIS unit elements (i.e., meta-atoms) in order to primarily investigate the performance of our DRL approach as the action space increases. The simulated BS precoding codebook comprised four unit-power beams, i.e., $\text{card}(\mathcal{V}) = 4$, which corresponds to the columns of the 2×2 DFT matrix. We allocated the first (last) two columns of this matrix as the available precoding vectors for the first (second) UE. Similar to other works in the field (e.g., [141] and [142]), we have also made the assumption that the meta-atoms of each RIS are controlled in groups of $N_{\text{group}} = 16$ so that all elements of the group are set with the same configuration. We will be referring to the total number of meta-atoms that can be independently controlled as N_{control} throughout this section (i.e., $N_{\text{control}} \triangleq N_{\text{tot}}/N_{\text{group}}$). It is noted that this implementation decision is expected to have a small adverse effect on the attained performance of all simulated methods; however, it vastly decreases the dimensionality of the combinatorial action space, rendering the ANN training, realized at the controller of the smart wireless environment, computationally tractable.

For both our simulated DRL-based orchestration controllers, namely, the neural ϵ-greedy and DQN algorithms, we have used a training period (i.e., number of i.i.d. channel realizations, each corresponding to a coherent time step) equal to the number of actions $\text{card}(\mathcal{A})$ multiplied by 50 in each trial. Their performance was then assessed by computing the average sum rate over 300 time steps, during which the controller acted with its deterministically greedy learned policy (i.e., without selecting a random action with probability ϵ). To avoid extreme initializations for the ANN parameters of the DRL approaches (see Algorithm 3) with the above evaluation strategy, we have performed averaging over five different trials. Note that, depending on the intended application, the training process might continue indefinitely; in that case, the controller needs to be evaluated at certain intervals using its exploration policy. For the baseline random and the optimal schemes, we have computed the average sum rates at the beginning of each trial using 300 i.i.d. channel realizations.

C. Sum-Rate Results Versus the Size of the Action Space

As discussed in Section IV-A, and particularly from the definition of the action space vector in (32), the action space increases exponentially with the number of the phase-tunable RIS meta-atoms. This implies that the number of observations required for the DRL-based approaches to associate observations to approximately optimal actions
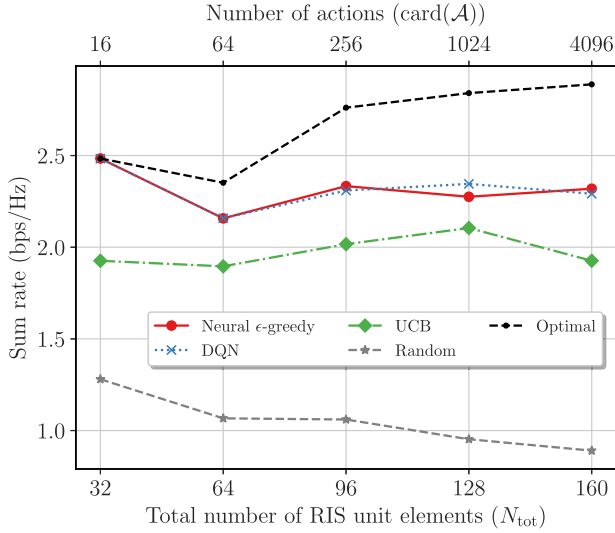
Fig. 7. *Average sum-rate performance in bps/Hz versus the total number of RIS elements, N_{tot}, for each of the $M = 2$ identical RISs and the size of the action space, card($\mathcal{A}$), achievable by the neural ϵ-greedy, DQN, and UCB algorithms. The sum-rate curves of the random selection and optimal schemes are also demonstrated.*

will ultimately lead to intractable training times. Consider a toy example, using the parameters in Table 4, with only one group of meta-atoms per RIS, i.e., $N_{tot} = 32$ elements, and 16 different actions, taking also the selection of the BS precoding matrix under consideration. Adding one phased-controllable group in each RIS, up to a total of ten groups, leads to 160 meta-atoms and action space, including 4096 choices. It is obvious that the action space can be extensively large even with RISs having meta-atoms with the lowest possible phase resolution $b = 1$.

The achievable average sum rates for the simulation setup described in Table 4, using all considered orchestration methods for solving $\mathcal{OP}_1$ without the constraint in (11c) for the individual rate requests, are depicted in Fig. 7 as functions of the number N_{tot} of the meta-atoms for each RIS and the size card($\mathcal{A}$) of the action space. The same results are also included in Fig. 8 in a normalized manner, by dividing each method's sum rate by the rate of the optimal scheme for all considered N_{tot} and card($\mathcal{A}$) values. As observed in both figures, all RL methods far outperform the random selection policy for all considered cases. It is also evident that both DRL methods are capable of maintaining sum rates higher than about 80% of the optimal rates in all cases, while being able to converge to the optimal RIS phase configurations in the simplest case of 32 meta-atoms per RIS. In fact, the proposed neural ϵ-greedy method and DQN exhibit identical performance, confirming our hypothesis: in smart wireless environments with i.i.d. channel realizations, bandit algorithms are akin to state-of-the-art MDP-based DRL approaches. Interestingly enough, the performance of our simple UCB method is capable to consistently outperform the baseline random selection policy while being reasonably close to both DRL

algorithms, even without relying on CSI observations. Nevertheless, a decreasing trend in the relative performance (w.r.t. the optimal sum rates) can be inferred from Fig. 8 with increasing values for N_{tot} and card($\mathcal{A}$), which we expect to continue in cases where the sizes of the RISs increase.

D. Sum-Rate Results Versus the Transmit Power

The observation and action spaces, defined in our DRL formulation in Section IV-A, are independent of the choice of the BS transmit power budget P. However, it is apparent from the SINR expression in (4) that the transmit power affects the reward values [see (37) and (38)] collected by the controller. Similar to Fig. 8, in Fig. 9, we demonstrate the normalized achievable sum-rate performance versus different P values and $N_{tot} = $ card($\mathcal{A}$) $= 64$ for the neural ϵ-greedy, DQN, and UCB approaches, as well as the random selection scheme. It is depicted that the considered DRL algorithms (i.e., the proposed neural ϵ-greedy and DQN) perform similarly, achieving around 75% of the optimum performance, nearly irrespective of the tested values for P. It is also shown that the performance of our UCB algorithm improves with increasing P, ranging from around 45% of the optimum sum rate at $P = 10$ dBm to around 70% when $P = 50$ dBm. Interestingly, for the latter large P value, our simple RL method attains an almost equal sum rate with both considered DRL methods. This implies that the average values of favorable actions become more easily discriminated from the expected rates of the on-average unfavorable ones. Finally, it can be observed that all considered (D)RL algorithms outperform the random selection scheme, whose normalized sum rate depends on
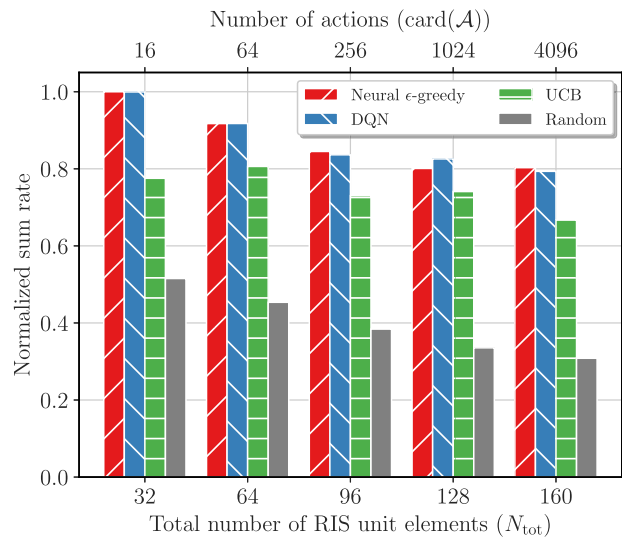
Fig. 8. *Average sum rates of the neural ϵ-greedy, DQN, and UCB algorithms, as well as of the random selection scheme illustrated in Fig. 7 when normalized over the sum-rate performance of the optimal scheme.*
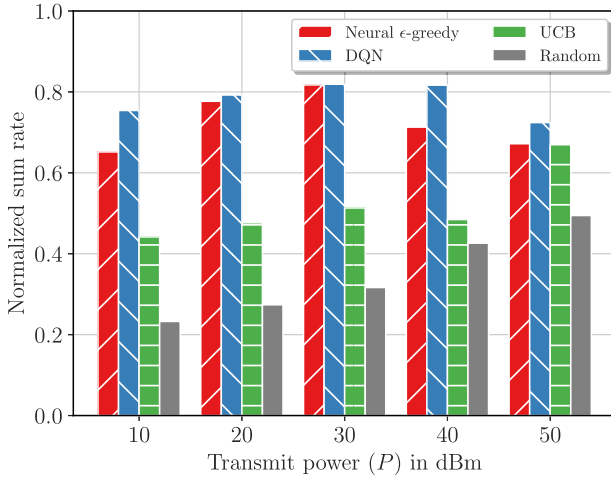
Fig. 9. *Average sum-rate performance in bps/Hz versus the transmit power P in dBm for the neural ε-greedy, DQN, and UCB algorithms, as well as the random selection scheme, normalized over the sum-rate performance of the optimal scheme. We have considered M = 2 identical RISs with each having N_{tot} = 32 meta-atoms, resulting in an action space of size card(A) = 64.*

the transmit power P but is always below the 45% of the optimum achievable sum-rate performance.

E. Reward Results Versus the Size of the Action Space

By solving $\mathcal{OP}_1$ with the constraint in (11c) for the per UE rate requests, Fig. 10 illustrates the average reward values obtained using (38) as functions of the N_{tot} for each RIS and card($\mathcal{A}$) similar to Fig. 7. In this figure, we used the same simulation parameters with Figs. 7 and 8 and set both R_1^{req} and R_2^{req} appearing in (11c) equal to 0.4 bps/Hz; this rate value corresponds to about half the optimal achievable rate when $N_{tot} = 32$. As clearly shown, the achieved reward with the random selection policy falls below zero for all investigated cases, indicating that the selected combinations of RIS phase profiles and BS precoders do not meet the UEs' rate requests. It is also demonstrated that the optimal policy attains lower average rewards than the achievable sum rates in Fig. 7, which witnesses that, at certain channel realizations, the UE rate requests cannot be satisfied. For this problem formulation, including the constraints (11c), it is interestingly shown in the figure that the performance of the proposed UCB method is at the same levels as that of the DRL methods, i.e., the proposed neural ε-greedy algorithm and DQN. We attribute this to the fact that the scale of the reward values is now larger due to the introduction of negative values in (38), thus facilitating the detection of actions of poor quality. More importantly, a decreasing trend is apparent on the performance of the considered methods, signifying that the respective orchestration mechanisms are less efficient than those in Fig. 7 in learning the environment. It can be finally concluded from Fig. 10, due to the positive average reward values, that all methods, expect the

baseline random scheme, are successful in satisfying the per UE rate constraints.

F. Orchestration With Partial CSI Observability

The observation of the evolving states of the wireless environment from the orchestration controller, via the collection of the CSI vectors in (33), is one of the key practical challenges with the proposed DRL formulation, as presented in Section V-B. The role of the controller is to associate the CSI observations to desirable RIS phase configurations. In practice, the associated computational cost and latency induced by the exchange of pilot signals for reliable CSI estimation are important considerations in deploying the proposed multi-RIS-empowered smart wireless environment. Therefore, it is essential for the envisioned DRL-based controller to rely on more efficient types of environmental observations.

In this section, we compare the performance of the considered orchestration methods between the original full CSI availability case and another case of partial CSI observability, according to which the wireless propagation environment includes strong LOS components, and the controller possesses the AoDs between each RIS and the UEs. The AoDs between the BS and each RIS are also assumed known due to the known placement of the RISs and the BS position. In the context of the latter partial CSI case, we do not deal with the problem of accurately estimating the angular channel parameters; instead, we assume computationally and sensingly autonomous RISs (e.g., [64], [102], [107], and [110]).

Up to this point (see Section VI-B with the simulation setup), static UEs with fixed positions have been considered. To properly evaluate the AoD-based partial
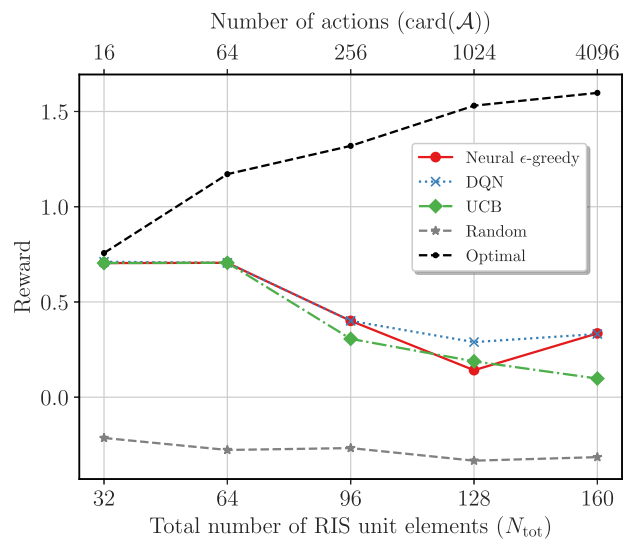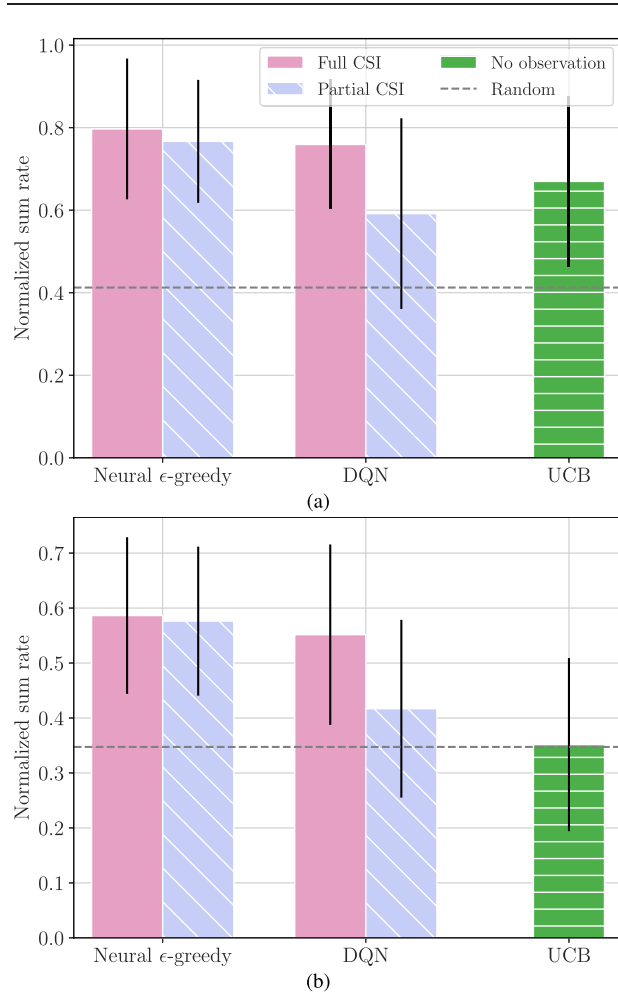

Fig. 10. *Average rewards using (38) for all proposed and benchmark orchestration methods versus the total number of RIS meta-atoms, N_{tot}, for each of the M = 2 identical RISs and the size of the action space, card(A).*

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

CSI observability case, we consider the following mobility scenario: two UEs are moving with a standard walking speed of 1.4 m/s in straight-line trajectories, angled at 45° and −45°, respectively, with respect to the x-axis. When their displacement from their starting positions exceeds 2 m, they turn around and move in the opposite directions. The channel coherence time in this setup is approximately 6 ms, which constitutes one time step in our DRL formulation. The rest of the simulation parameters and the algorithmic settings remain as described in Sections VI-A–VI-C, with the following exception regarding the structure of the ANNs employed by the DRL algorithms. Since the dimension of the observation space is small enough, the two convolutional and max-pooling layers are not needed. Instead, they are all replaced by one fully connected layer of 32 units. Since the sum rates depended on the positions of the UEs, we evaluated the considered algorithms

Table 5 Execution Time of the Considered Methods in Time Steps per Second

N_{tot}	Neural ϵ-greedy	DQN	UCB	Optimal	Random
32	72	45	2603	493	3012
64	61	33	2100	112	2489
96	58	31	1532	28	1744
128	55	25	726	6	661
160	48	24	651	1	188

(including the benchmark schemes) at 15 evenly spaced intervals during the training process, each consisting of 300 time steps (i.e., CSI realizations), and we considered the average achievable sum-rate performance across all intervals.

The comparison of the sum rates between the full and partial CSI observation cases when considering the DRL methods neural ϵ-greedy and DQN is illustrated in Fig. 11 for the setups with $N_{tot} = \{64, 128\}$ total numbers of RIS unit elements. The performance of the proposed UCB method is also included, which constitutes an RL algorithm that does not rely on any type of environmental observation. More specifically, the individual rewards serve as feedback on the selected actions. As depicted in the figure, the normalized performances of the proposed neural ϵ-greedy and the DQN algorithms are similar when CSI observations are involved. Interestingly, the former is also shown to perform equally well when the observations consist of AoDs. Conversely, DQN exhibits a noticeable decrease in performance when using partial, instead of full CSI, observations. It is finally shown that the normalized performance of the UCB method decreases more rapidly, compared to the DRL methods, as N_{tot} increases.

Fig. 11. *Average sum rates of the neural ϵ-greedy, DQN, and UCB algorithms, as well as of the random selection scheme, when normalized versus the sum-rate performance of the optimal scheme. The DRL algorithms have been evaluated for both the full and partial CSI observation cases, considering a scenario with two mobile UEs and two values for the total number of RIS meta-atoms, N_{tot}. The vertical error bars in the figure indicate the respective standard deviations. (a) $N_{tot} = 64$. (b) $N_{tot} = 128$.*

G. Methods' Execution Time Comparison

To complement the asymptotic time complexities of the DRL algorithms discussed in Section III-E, we herein quantify the execution time of each considered method in the simulation setup presented in Section VI-B. For this evaluation, we have used a desktop computer with an Intel i5-8400 processor, 16 GB of RAM, and an NVIDIA GTX-1060 with a 6-GB VRAM, and the DRL algorithms were implemented in Tensorflow.[11] The mean execution time in time steps per second for all simulated methods is given in Table 5. As it can be concluded from the comparison of the two DRL algorithms, DQN is slower than neural ϵ-greedy since the former adopts an additional (second) neural network. However, both algorithms exhibit performance metrics of the same order of magnitude. It is also shown that the UCB method is considerably faster, rivaling the performance of the baseline random selection scheme. Evidently, the optimal scheme (i.e., exhaustive search) runs faster than both DRL methods in the setups with the smaller action spaces although it becomes prohibitively expensive as the sizes of the RISs increase. A further

---

[11]The code for the performance results presented in this article is available at https://github.com/NoesysLab/DRL_RIS_Tutorial

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

interesting result that can be inferred from the table is that the execution time of both DRL methods does not increase significantly for large numbers N_{tot} of the RISs' meta-atoms. This behavior is justified by the parallel computing capabilities of the utilized GPU.

VII. CONCLUSION

In this article, focusing on the emerging RIS technology for programmable propagation of information-bearing signals in the era of beyond 5G wireless communications, we studied the dynamic orchestration of the promising, but challenging, multi-RIS-empowered smart wireless environments by means of state-of-the-art DRL methods. Our comprehensive treatment included a description of a generic model for such communication systems, which encompasses various crucial aspects of reconfigurable radio wave propagation. A thorough theoretical introduction of the RL area was provided to illuminate the basic principles, categorization, and applicability scenarios of the most prominent DRL algorithms with emphasis on those deployed in the wireless communications field. We then described the exemplar sum-rate maximization problem via a quite general DRL formulation that aligns itself with the existing literature. Relevant works were discussed and taxonomized in order to give a clear overview of the adoption of the DRL approaches designed by the community working on RISs and RIS-empowered smart wireless environments. The literature overview has been accompanied by a

discussion on fundamental challenges, key opportunities, and important future directions of this evolving research area.

Apart from the survey intention of this article, our technical contribution lies in the alternative treatment of the online sum-rate maximization objective as a multiarmed bandits' problem, instead of the MDP-inspired DRL formulations typically found in the field. The proposed framework leveraged the fact that, at any time instant, the design parameters of the multi-RIS-enabled smart wireless environment (i.e., the RIS phase profiles and the BS precoding matrix) have no effect on the state of the environment (i.e., the channel gain matrices) of the next time instant. Our extensive numerical evaluations showcased that the proposed multiarmed bandits' orchestration controller is capable of attaining increased performance in a number of setups. Interestingly, this achievable performance is similar to DQN-based orchestration but with simpler implementation and, consequently, faster execution time. We also presented a simple UCB strategy, selecting the action that has the highest upper bound estimate of the average gain associated with it, which did not require CSI observability, but relied on the feedback of SINR measurements for the reward computation. It was demonstrated that this simple orchestration scheme performs sufficiently close to the considered DRL algorithms in certain scenarios while always outperforming the baseline random selection scheme. ∎

REFERENCES

[1] "3GPP releases 16 & 17 & beyond," 5G Americas, Bellevue, WA, USA, White Paper, Jan. 2021.

[2] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[3] 3GPP. (2022). *Release 18*. [Online]. Available: https://www.3gpp.org/release18

[4] "The next hyper—Connected experience for all," Samsung 6G Vis., Suwon-si, South Korea, White Paper, Jun. 2020.

[5] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, Oct. 2020.

[6] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, pp. 1–17, May 2021.

[7] "Cisco annual internet report (2018–2023)," Cisco, San Jose, CA, USA, White Paper C11-741490-01, Mar. 2020.

[8] "Accelerating the adoption of AI in programmable 5G networks," Ericsson, Stockholm, Sweden, White Paper, Jul. 2021.

[9] "AI and ML—Enablers for beyond 5G networks," 5G PPP Technol. Board, White Paper, May 2021, doi: 10.5281/zenodo.4299895.

[10] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.

[11] G. C. Alexandropoulos *et al.*, "Time reversal for 6G wireless communications: Novel experiments, opportunities, and challenges," *IEEE Veh. Technol. Mag.*, to be published. [Online]. Available: https://arxiv.org/abs/2206.08419

[12] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun.*

Mag., vol. 56, no. 9, pp. 162–169, Sep. 2018.

[13] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, May 2019.

[14] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[15] E. Basar, M. D. Renzo, J. D. Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[16] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Nov. 2020.

[17] C. Huang *et al.*, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[18] G. C. Alexandropoulos, G. Lerosey, M. Debbah, and M. Fink, "Reconfigurable intelligent surfaces and metamaterials: The potential of wave propagation control for 6G wireless communications," *IEEE ComSoc TCCN Newslett.*, vol. 6, no. 1, pp. 25–37, Jun. 2020.

[19] E. C. Strinati *et al.*, "Wireless environment as a service enabled by reconfigurable intelligent surfaces: The RISE-6G perspective," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2021, pp. 562–567.

[20] G. C. Alexandropoulos, N. Shlezinger, and P. del Hougne, "Reconfigurable intelligent surfaces for rich scattering wireless communications: Recent experiments, challenges, and opportunities," *IEEE*

Commun. Mag., vol. 59, no. 6, pp. 28–34, Jun. 2021.

[21] E. C. Strinati *et al.*, "Reconfigurable, intelligent, and sustainable wireless environments for 6G smart connectivity," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 99–105, Oct. 2021.

[22] Q. Wu *et al.*, "Intelligent reflecting surface aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[23] Q.-U.-A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface-assisted multi-user MISO communication: Channel estimation and beamforming design," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 661–680, 2020.

[24] H. Liu, X. Yuan, and Y.-J.-A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, Nov. 2020.

[25] L. Yang, F. Meng, Q. Wu, D. B. da Costa, and M.-S. Alouini, "Accurate closed-form approximations to channel distributions of RIS-aided wireless systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1985–1989, Nov. 2020.

[26] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, F. Chen, and S. Mumtaz, "Adaptive transmission for reconfigurable intelligent surface-assisted OFDM wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2653–2665, Nov. 2020.

[27] L. Wei, C. Huang, G. C. Alexandropoulos, and C. Yuen, "Parallel factor decomposition channel estimation in RIS-assisted multi-user MISO communication," in *Proc. IEEE 11th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jun. 2020, pp. 1–5.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

[28] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[29] B. Deepak, R. S. P. Sankar, and S. P. Chepuri, "Channel estimation for RIS-assisted millimeter-wave MIMO systems," 2020, *arXiv:2011.00900*.

[30] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, M. D. Renzo, and F. Chen, "Reconfigurable intelligent surfaces with reflection pattern modulation: Beamforming design and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 741–754, Feb. 2021.

[31] M. Guo and M. C. Gursoy, "Channel estimation for intelligent reflecting surface assisted wireless communications," 2021, *arXiv:2104.01221*.

[32] G. T. de Araujo, A. L. F. de Almeida, and R. Boyer, "Channel estimation for intelligent reflecting surface assisted MIMO systems: A tensor modeling approach," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 789–802, Apr. 2021.

[33] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. D. Renzo, and B. Ottersten, "Reconfigurable intelligent surface-assisted cell-free massive MIMO systems over spatially-correlated channels," 2021, *arXiv:2104.08648*.

[34] X. Chen, J. Shi, Z. Yang, and L. Wu, "Low-complexity channel estimation for intelligent reflecting surface-enhanced massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 996–1000, May 2021.

[35] X. Guan, Q. Wu, and R. Zhang, "Anchor-assisted channel estimation for intelligent reflecting surface aided multiuser communication," 2021, *arXiv:2102.10886*.

[36] H. Guo and V. K. N. Lau, "Cascaded channel estimation for intelligent reflecting surface assisted multiuser MISO systems," 2021, *arXiv:2108.09002*.

[37] X. Shi, J. Wang, G. Chen, and J. Song, "Triple-structured compressive sensing-based channel estimation for RIS-aided MU-MIMO systems," 2021, *arXiv:2108.13765*.

[38] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar, and D. R. Smith, "Dynamic metasurface antennas for 6G extreme massive MIMO communications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 106—113, Apr. 2021.

[39] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[40] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4659–4663.

[41] T. L. Jensen and E. D. Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 5000–5004.

[42] J.-M. Kang, "Intelligent reflecting surface: Joint optimal training sequence and refection pattern," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1784–1788, Aug. 2020.

[43] S. Xia and Y. Shi, "Intelligent reflecting surface for massive device connectivity: Joint activity detection and channel estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 5175–5179.

[44] G. T. de Araujo and A. L. F. de Almeida, "PARAFAC-based channel estimation for intelligent reflective surface assisted MIMO system," in *Proc. IEEE 11th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jun. 2020, pp. 1–5.

[45] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2020.

[46] Z. Wan, Z. Gao, F. Gao, M. D. Renzo, and M.-S. Alouini, "Terahertz massive MIMO with holographic reconfigurable intelligent surfaces," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4732–4750, Jul. 2021.

[47] J. Yuan, G. C. Alexandropoulos, E. Kofidis, T. L. Jensen, and E. D. Carvalho, "Tensor-based channel tracking for RIS-empowered multi-user MIMO wireless systems," 2022, *arXiv:2202.08315*.

[48] X. Ying, U. Demirhan, and A. Alkhateeb, "Relay aided intelligent reconfigurable surfaces: Achieving the potential without so many antennas," 2020, *arXiv:2006.06644*.

[49] E. Bjornson, O. Ozdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

[50] M. Nemati, J. Park, and J. Choi, "RIS-assisted coverage enhancement in millimeter-wave cellular networks," *IEEE Access*, vol. 8, pp. 188171–188185, 2020.

[51] L. Yang, Y. Yang, M. O. Hasna, and M.-S. Alouini, "Coverage, probability of SNR gain, and DOR analysis of RIS-aided communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1268–1272, Aug. 2020.

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

[53] I. Yildirim, F. Kilinc, E. Basar, and G. C. Alexandropoulos, "Hybrid RIS-empowered reflection and decode-and-forward relaying for coverage extension," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1692–1696, May 2021.

[54] T. Ma, Y. Xiao, X. Lei, W. Xiong, and Y. Ding, "Indoor localization with reconfigurable intelligent surface," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 161–165, Jan. 2021.

[55] H. Zhang *et al.*, "MetaRadar: Indoor localization by reconfigurable metamaterials," 2020, *arXiv:2008.02459*.

[56] A. Elzanaty, A. Guerra, F. Guidi, and M.-S. Alouini, "Reconfigurable intelligent surfaces for localization: Position and orientation error bounds," 2020, *arXiv:2009.02818*.

[57] K. Keykhosravi, M. F. Keskin, G. Seco-Granados, and H. Wymeersch, "SISO RIS-enabled joint 3D downlink localization and synchronization," 2020, *arXiv:2011.02391*.

[58] S. Buzzi, E. Grossi, M. Lops, and L. Venturino, "Radar target detection aided by reconfigurable intelligent surfaces," 2021, *arXiv:2104.00768*.

[59] M. Rahal, B. Denis, K. Keykhosravi, B. Uguen, and H. Wymeersch, "RIS-enabled localization continuity under near-field conditions," 2021, *arXiv:2109.11965*.

[60] Z. Abu-Shaban, K. Keykhosravi, M. F. Keskin, G. C. Alexandropoulos, G. Seco-Granados, and H. Wymeersch, "Near-field localization with a reconfigurable intelligent surface acting as lens," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[61] Z. Yang, H. Zhang, B. Di, H. Zhang, K. Bian, and L. Song, "Wireless indoor simultaneous localization and mapping using reconfigurable intelligent surface," 2021, *arXiv:2107.01582*.

[62] R. S. P. Sankar, B. Deepak, and S. P. Chepuri, "Joint communication and radar sensing with reconfigurable intelligent surfaces," 2021, *arXiv:2105.01966*.

[63] C. L. Nguyen, O. Georgiou, and G. Gradoni, "Reconfigurable intelligent surfaces and machine learning for wireless fingerprinting localization," 2020, *arXiv:2010.03251*.

[64] G. C. Alexandropoulos, I. Vinieratou, and H. Wymeersch, "Localization via multiple reconfigurable intelligent surfaces equipped with single receive RF chains," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 1072–1076, May 2022.

[65] K. Keykhosravi, G. Seco-Granados, G. C. Alexandropoulos, and H. Wymeersch, "RIS-enabled self-localization: Leveraging

controllable reflections with zero access points," in *Proc. IEEE ICC*. Seoul South Korea, Jun. 2022, pp. 1–6.

[66] J. Chen, Y.-C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82599–82612, 2019.

[67] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

[68] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[69] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for secure IRS-assisted multiuser MISO systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2019, pp. 1–6.

[70] X. Yu, D. Xu, and R. Schober, "Enabling secure wireless communications via intelligent reflecting surfaces," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[71] A. Almohamad *et al.*, "Smart and secure wireless communications via reflecting intelligent surfaces: A short survey," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1442–1456, 2020.

[72] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 108–112, Jan. 2020.

[73] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7851–7866, Dec. 2020.

[74] L. Dong and H.-M. Wang, "Enhancing secure MIMO transmission via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7543–7556, Nov. 2020.

[75] F. Shu *et al.*, "Enhanced secrecy rate maximization for directional modulation networks via IRS," 2020, *arXiv:2008.05067*.

[76] G. C. Alexandropoulos, K. Katsanos, M. Wen, and D. B. D. Costa, "Safeguarding MIMO communications with reconfigurable metasurfaces and artificial noise," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[77] M. AL-Mekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," 2021, *arXiv:2108.02346*.

[78] M. Jung, W. Saad, M. Debbah, and C. S. Hong, "On the optimality of reconfigurable intelligent surfaces (RISs): Passive beamforming, modulation, and resource allocation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4347–4363, Jul. 2021.

[79] A. L. Moustakas, G. C. Alexandropoulos, and M. Debbah, "Capacity optimization using reconfigurable intelligent surfaces: A large system approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[80] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

[81] L. You *et al.*, "Reconfigurable intelligent surfaces-assisted multiuser MIMO uplink transmission with partial CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5613–5627, Sep. 2021.

[82] D. Xu, X. Yu, D. W. K. Ng, and R. Schober, "Resource allocation for active IRS-assisted multiuser communication systems," 2021, *arXiv:2108.13033*.

[83] Z. Yang *et al.*, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 665–679, Jan. 2022.

[84] G. Yang, Y. Liao, Y.-C. Liang, O. Tirkkonen, G. Wang, and X. Zhu, "Reconfigurable intelligent

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

surface empowered device-to-device communication underlaying cellular networks," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7790–7805, Nov. 2021.

[85] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[86] A. Zappone, M. D. Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[87] C. Xiang, S. Zhang, S. Xu, and G. C. Alexandropoulos, "Self-calibrating indoor localization with crowdsourcing fingerprints and transfer learning," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[88] S. BelMannoubi and H. Touati, "Deep neural networks for indoor localization using WiFi fingerprints," in *Proc. MSPN*. Mohammedia, Morocco, Apr. 2019, pp. 247–258.

[89] C. Xiang et al., "Robust sub-meter level indoor localization with a single WiFi access point—Regression versus classification," *IEEE Access*, vol. 7, pp. 146309–146321, 2019.

[90] C. Huang, G. C. Alexandropoulos, A. Zappone, C. Yuen, and M. Debbah, "Deep learning for UL/DL channel calibration in generic massive MIMO systems," in *Proc. IEEE ICC*. Shanghai, China, May 2019, pp. 1–6.

[91] Q. Hu, F. Gao, H. Zhang, S. Jin, and G. Y. Li, "Deep learning for channel estimation: Interpretation, performance, and comparison," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2398–2412, Apr. 2021.

[92] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, 2018.

[93] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Tech.*, vol. 68, no. 3, pp. 3027–3032, Jan. 2019.

[94] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 800–805.

[95] N. N. Sapavath, S. Safavat, and D. B. Rawat, "On the machine learning–based smart beamforming for wireless virtualization with large-scale MIMO system," *Trans. Emerg. Telecommun. Technol.*, vol. 30, no. 9, Aug. 2019, Art. no. e3713.

[96] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[97] B. Yang, X. Cao, C. Huang, C. Yuen, L. Qian, and M. D. Renzo, "Intelligent spectrum learning for wireless networks with reconfigurable intelligent surfaces," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3920–3925, Apr. 2021.

[98] B. Sheen, J. Yang, X. Feng, and M. M. U. Chowdhury, "A deep learning based modeling of reconfigurable intelligent surface assisted wireless communications for phase shift configuration," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 262–272, 2021.

[99] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, "Unsupervised learning for passive beamforming," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1052–1056, May 2020.

[100] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "An interpretable neural network for configuring programmable wireless environments," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[101] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE Access*, vol. 9, pp. 44304–44321, 2021.

[102] G. C. Alexandropoulos and E. Vlachos, "A hardware architecture for reconfigurable

intelligent surfaces with minimal active elements for explicit channel estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 9175–9179.

[103] C. Huang, R. Mo, and Y. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Jun. 2020.

[104] A. M. Elbir, A. Papazafeiropoulos, P. Kourtessis, and S. Chatzinotas, "Deep channel learning for large intelligent surfaces aided mm-Wave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1447–1451, Sep. 2020.

[105] S. Khan, K. S Khan, N. Haider, and S. Y. Shin, "Deep-learning-aided detection for reconfigurable intelligent surfaces," 2019, *arXiv:1910.09136*.

[106] Q. Ma, G. D. Bai, H. B. Jing, C. Yang, L. L. Li, and T. J. Cui, "Smart metasurface with self-adaptively reprogrammable functions," *Light, Sci. Appl.*, vol. 8, p. 98, Oct. 2019.

[107] C. Liaskos et al., "ABSense: Sensing electromagnetic waves on metasurfaces via ambient compilation of full absorption," 2019, *arXiv:1907.04811*.

[108] Q. Ma et al., "Smart sensing metasurface with self-defined functions in dual polarizations," *Nanophotonics*, vol. 9, no. 10, pp. 3271–3278, Aug. 2020.

[109] G. C. Alexandropoulos, N. Shlezinger, I. Alamzadeh, M. F. Imani, H. Zhang, and Y. C. Eldar, "Hybrid reconfigurable intelligent metasurfaces: Enabling simultaneous tunable reflections and sensing for 6G wireless communications," 2021, *arXiv:2104.04690*.

[110] I. Alamzadeh, G. C. Alexandropoulos, N. Shlezinger, and M. F. Imani, "A reconfigurable intelligent surface with integrated sensing capability," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Oct. 2021.

[111] H. Zhang, N. Shlezinger, I. Alamzadeh, G. C. Alexandropoulos, M. F. Imani, and Y. C. Eldar, "Channel estimation with simultaneous reflecting and sensing reconfigurable intelligent metasurfaces," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 536–540.

[112] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

[113] L. Li et al., "Machine-learning reprogrammable metasurface imager," *Nature Commun.*, vol. 10, no. 1, p. 1082, Jun. 2019.

[114] X. Cao et al., "AI-assisted MAC for reconfigurable intelligent-surface-aided wireless networks: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 21–27, Jun. 2021.

[115] L. Wang, N. Shlezinger, G. C. Alexandropoulos, H. Zhang, B. Wang, and Y. C. Eldar, "Jointly learned symbol detection and signal reflection in RIS-aided multi-user MIMO systems," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2021, pp. 715–721.

[116] K. Stylianopoulos, N. Shlezinger, P. D. Hougne, and G. C. Alexandropoulos, "Deep-learning-assisted configuration of reconfigurable intelligent surfaces in dynamic rich-scattering environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8822–8826.

[117] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light, Sci. Appl.*, vol. 3, no. 10, p. e218, Oct. 2014.

[118] M. Jian et al., "Reconfigurable intelligent surfaces for wireless communications: Overview of hardware designs, channel models, and estimation techniques," *Intell. Converged Netw.*, vol. 3, no. 1, pp. 1–32, Mar. 2022.

[119] W. Tang et al., "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," 2019,

arXiv:1911.05326.

[120] L. Dai et al., "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45913–45923, 2020.

[121] B. O. Zhu, J. Zhao, and Y. Feng, "Active impedance metasurface with full 360° reflection phase tuning," *Sci. Rep.*, vol. 3, pp. 3059–3064, Oct. 2013.

[122] F. Liu et al., "Intelligent metasurfaces with continuously tunable local surface impedance for multiple reconfigurable functions," *Phys. Rev. A, Gen. Phys.*, vol. 11, no. 4, Apr. 2019, Art. no. 044024.

[123] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[124] L. Pulido-Mancera, P. T. Bowen, M. F. Imani, N. Kundtz, and D. Smith, "Polarizability extraction of complementary metamaterial elements in waveguides for aperture modeling," *Phys. Rev. B, Condens. Matter*, vol. 96, no. 23, Dec. 2017, Art. no. 235402.

[125] A. Clemente, L. Dussopt, R. Sauleau, P. Potier, and P. Pouliguen, "1-bit reconfigurable unit cell based on PIN diodes for transmit-array applications in *X*-band," *IEEE Trans. Antennas Propag.*, vol. 60, no. 5, pp. 2260–2269, May 2012.

[126] *5G; NR; Physical Layer Procedures for Data*, Standard 38.214, Version 16.2.0, 3GPP, Tech. Rep., Jul. 2020.

[127] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.

[128] G. C. Alexandropoulos, M. Kountouris, and I. Atzeni, "User scheduling and optimal power allocation for full-duplex cellular networks," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2016, pp. 1–6.

[129] A. Ghazanfari, H. V. Cheng, E. Bjornson, and E. G. Larsson, "Enhanced fairness and scalability of power control schemes in multi-cell massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2878–2890, May 2020.

[130] R. Faqiri, C. Saigre-Tardif, G. C. Alexandropoulos, N. Shlezinger, M. F. Imani, and P. del Hougne, "PhysFad: Physics-based end-to-end channel modeling of RIS-parametrized environments with adjustable fading," 2022, *arXiv:2202.02673*.

[131] G. C. Alexandropoulos, V. I. Barousis, and C. B. Papadias, "Precoding for multiuser MIMO systems with single-fed parasitic antenna arrays," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 3897–3902.

[132] G. Gradoni and M. D. Renzo, "End-to-end mutual coupling aware communication model for reconfigurable intelligent surfaces: An electromagnetic-compliant approach based on mutual impedances," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 938–942, May 2021.

[133] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels—A survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.

[134] J. He, T. Kim, H. Ghauch, K. Liu, and G. Wang, "Millimeter wave MIMO channel tracking systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 8–12.

[135] G. C. Alexandropoulos and S. Chouvardas, "Low complexity channel estimation for millimeter wave systems with hybrid A/D antenna processing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[136] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," 2018, *arXiv:1804.01908*.

[137] G. C. Alexandropoulos and C. B. Papadias, "A reconfigurable iterative algorithm for the K-user MIMO interference channel," *Signal Process.*, vol. 93, no. 12, pp. 3353–3362, Dec. 2013.

[138] F. R. Farrokhi, A. Lozano, G. J. Foschini, and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

R. A. Valenzuela, "Spectral efficiency of FDMA/TDMA wireless systems with transmit and receive antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 591–599, Oct. 2002.

[139] G. C. Alexandropoulos, I. Vinieratou, M. Rebato, L. Rose, and M. Zorzi, "Uplink beam management for millimeter wave cellular MIMO systems with hybrid beamforming," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–7.

[140] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[141] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, "Energy efficient multi-user MISO communication using low resolution large intelligent surfaces," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[142] G. C. Alexandropoulos, S. Samarakoon, M. Bennis, and M. Debbah, "Phase configuration learning in wireless networks with multiple reconfigurable intelligent surfaces," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2020, pp. 1–6.

[143] H. Gacanin and M. D. Renzo, "Wireless 2.0: Toward an intelligent radio environment empowered by reconfigurable meta-surfaces and artificial intelligence," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 74–82, Dec. 2020.

[144] A. M. Elbir and K. V. Mishra, "A survey of deep learning architectures for intelligent reflecting surfaces," 2020, *arXiv:2009.02540*.

[145] S. Zhang, M. Li, M. Jian, Y. Zhao, and F. Gao, "AIRIS: Artificial intelligence enhanced signal processing in reconfigurable intelligent surface communications," 2021, *arXiv:2106.00171*.

[146] R. S. Sutton and A. G. Barto, *Reinforcement Learning I: Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[147] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Nashua, NH, USA: Athena Scientific, 2000.

[148] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Proc. NeurIPS*, 2018, pp. 2455–2467.

[149] A. Lazaridis, A. Fachantidis, and I. Vlahavas, "Deep reinforcement learning: A state-of-the-art walkthrough," *J. Artif. Intell. Res.*, vol. 69, pp. 1421–1471, Dec. 2020.

[150] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[151] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[152] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*.

[153] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Proc. PMLR*, vol. 120, Jun. 2020, pp. 486–489.

[154] R. S. Sutton, D. McAllester, S. Singh, and

Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. NIPS*, Dec. 1999, pp. 1057–1063.

[155] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, 1992.

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

[157] C. Berner *et al.*, "Dota 2 with large scale deep reinforcement learning," 2019, *arXiv:1912.06680*.

[158] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[159] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 661–670.

[160] C. Riquelme, G. Tucker, and J. Snoek, "Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling," in *Proc. Int. Conf. Learn. Represent.* Vancouver, BC, Canada, May 2018.

[161] X. Gao, Y. Liu, X. Liu, and L. Song, "Machine learning empowered resource allocation in IRS aided MISO-NOMA networks," 2021, *arXiv:2103.11791*.

[162] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[163] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[164] J. Hu, H. Zhang, K. Bian, M. D. Renzo, Z. Han, and L. Song, "MetaSensing: Intelligent metasurface assisted RF 3D sensing by deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2182–2197, Jul. 2021.

[165] K. K. Nguyen, S. Khosravirad, D. B. da Costa, L. D. Nguyen, and T. Q. Duong, "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning," 2021, *arXiv:2105.14142*.

[166] K. Feng, Q. Wang, X. Li, and C. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[167] C. Huang *et al.*, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[168] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable

intelligent surfaces," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[169] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1057–1071, Apr. 2021.

[170] J. Kim, S. Hosseinalipour, T. Kim, D. J. Love, and C. G. Brinton, "Multi-IRS-assisted multi-cell uplink MIMO communications under imperfect CSI: A deep reinforcement learning approach," 2020, *arXiv:2011.01141*.

[171] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, and S. Sharafeddine, "Reconfigurable intelligent surface enabled vehicular communication: Joint user scheduling and passive beamforming," 2021, *arXiv:2101.12247*.

[172] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[173] Q. Zhang, W. Saad, and M. Bennis, "Distributional reinforcement learning for mmWave communications with intelligent reflectors on a UAV," 2020, *arXiv:2011.01840*.

[174] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[175] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proc. ECML*. Berlin, Germany, 2005, pp. 437–448.

[176] M. Collier and H. U. Llorens, "Deep contextual multi-armed bandits," 2018, *arXiv:1807.09809*.

[177] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," in *Proc. NeurIPS*, vol. 32, 2019, pp. 8353–8365.

[178] G. C. Alexandropoulos and M. Kountouris, "Maximal ratio transmission in wireless Poisson networks under spatially correlated fading channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 1–6.

[179] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. AAAI Conf. Artificial Intell.*, vol. 32, no. 1. New Orleans, LA, USA, 2018, pp. 2892–2901.

[180] I. Akkaya *et al.*, "Solving Rubik's cube with a robot hand," 2019, *arXiv:1910.07113*.

[181] A. Tavakoli, F. Pardo, and P. Kormushev, "Action branching architectures for deep reinforcement learning," in *Proc. AAAI Conf. Artificial Intell.*, vol. 32, no. 1. New Orleans, LA, USA, Feb. 2018.

[182] G. Dulac-Arnold *et al.*, "Deep reinforcement learning in large discrete action spaces," 2015, *arXiv:1512.07679*.

[183] N. Yoshida, "Q-networks for binary vector actions," 2015, *arXiv:1512.01332*.

ABOUT THE AUTHORS

George C. Alexandropoulos (Senior Member, IEEE) received the Engineering Diploma, M.A.Sc., and Ph.D. degrees in computer engineering and informatics from the School of Engineering, University of Patras, Patras, Greece, in 2003, 2005, and 2010, respectively.

He has held research positions at various Greek universities and research institutes, as well as at the Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei Technologies France, Boulogne-Billancourt, France. He is currently an Assistant Professor with the Department of Informatics and Telecommunications, School of Sciences, National and Kapodistrian University of Athens (NKUA), Athens, Greece. He is also a Principal Researcher with the Technology Innovation Institute, Abu Dhabi, United Arab Emirates.

His research interests span the general areas of algorithmic design and performance analysis for wireless networks with an emphasis on multiantenna transceiver hardware architectures, active and passive reconfigurable metasurfaces, integrated communications and sensing, millimeter-wave and THz communications, and distributed machine learning algorithms.

Prof. Alexandropoulos is also a Senior Member of the IEEE Communications, Signal Processing, and Information Theory Societies and a registered Professional Engineer of the Technical Chamber of Greece. He received the Best Ph.D. Thesis Award in 2010, the IEEE Communications Society Best Young Professional in Industry Award 2018, the EURASIP Best Paper Award of the *Journal on Wireless Communications and Networking* in 2021, the IEEE Marconi Prize Paper Award in Wireless Communications in 2021, and a Best Paper Award from the IEEE Global Communications Conference (GLOBECOM) 2021. In the past, he has held various

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Alexandropoulos *et al.*: Pervasive Machine Learning for Smart Radio Environments Enabled by RISs

fixed-term and guest editorial positions for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE COMMUNICATIONS LETTERS and for various special issues at IEEE journals. He also serves as an Editor for IEEE WIRELESS COMMUNICATIONS LETTERS, *Computer Networks* (Elsevier), *Frontiers in Communications and Networks*, and the *ITU Journal on Future and Evolving Technologies*. He is also a Distinguished Lecturer of the IEEE Communications Society. He has participated and/or technically managed more than ten European Union (EU) research and innovation projects, as well as several Greek and international research projects. He is also a NKUA's Principal Investigator for the EU H2020 RISE-6G Research and Innovation Project dealing with RIS-empowered smart wireless environments. More information is available at www.alexandropoulos.info.

Kyriakos Stylianopoulos (Graduate Student Member, IEEE) graduated from the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece, in 2019. He received the M.Phil. degree from the University of Cambridge, Cambridge, U.K., in 2020. He is currently working toward the Ph.D. degree at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, working on the intersection of machine learning and beyond 5G wireless communications.

His broader research interests include pattern recognition, deep learning, reinforcement learning, observational astronomy, reconfigurable intelligent surfaces and smart radio environments, and next-generation wireless communication networks.

Chongwen Huang (Member, IEEE) received the B.Sc. degree from Nankai University, Tianjin, China, in 2010, the M.Sc. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2013, and the Ph.D. degree from the Singapore University of Technology and Design, Singapore, in 2019.

He joined the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, as a Research Engineer, in 2013. Since 2020, he has been a tenure-track Young Professor with Zhejiang University, Hangzhou, China. His main research interests are focused on holographic multiple-input–multiple-output (MIMO) surface/reconfigurable intelligent surfaces, beyond 5th generation (B5G)/sixth-generation (6G) wireless communications, mmWave/THz communications, and deep learning technologies for wireless communications.

Dr. Huang was a recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2021, the Singapore Government Ph.D. Scholarship, and the PHC Merlion Ph.D. Grant 2016–2019 for studying in CentraleSupélec, France. He has served as the Chair of some wireless communications flagship conferences, including 2021 IEEE Wireless Communications and Networking Conference (WCNC), 2021 IEEE Vehicular Technology Conference (VTC-Fall), and IEEE Wireless Communications and Signal Processing (WCSP) 2021. He is also an Editor for seven journals, including the IEEE COMMUNICATIONS LETTERS and *Signal Processing* (Elsevier).

Chau Yuen (Fellow, IEEE) received the B.Eng. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively.

He was a Postdoctoral Fellow with Lucent Technologies Bell Labs, Murray Hill, NJ, USA, in 2005, and a Visiting Assistant Professor with The Hong Kong Polytechnic University, Hong Kong, in 2008. From 2006 to 2010, he was with the Institute for Infocomm Research, Singapore, where he was involved in an industrial project on developing an 802.11n Wireless LAN system and participated actively in 3GPP LTE and LTE-Advanced standardization. Since 2010, he has been with the Singapore University of Technology and Design, Singapore.

Dr. Yuen received the IEEE Marconi Prize Paper Award in Wireless Communications and the EURASIP Best Paper Award for *Journal on Wireless Communications and Networking* in 2021, the IEEE Asia Pacific Outstanding Young Researcher Award in 2012, and the IEEE VTS Singapore Chapter Outstanding Service Award in 2019. He serves as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE SYSTEMS JOURNAL.

Mehdi Bennis (Fellow, IEEE) is currently a tenured Full Professor with the Centre for Wireless Communications, University of Oulu, Oulu, Finland, the Academy of Finland Research Fellow, and the Head of the Intelligent Connectivity and Networks/Systems Group (ICON), Oulu. He has published more than 200 research papers in international conferences, journals, and book chapters. His main research interests are in radio resource management, heterogeneous networks, game theory, and distributed machine learning in 5G networks and beyond.

Dr. Bennis was a recipient of several prestigious awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best Paper Award for the *Journal on Wireless Communications and Networking*, the All-University of Oulu Award for Research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, and the 2020 Clarviate Highly Cited Researcher by the Web of Science.

Mérouane Debbah (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the École Normale Supérieure Paris-Saclay, Cachan, France, in 1999 and 2002, respectively.

He was with Motorola Labs, Saclay, France, from 1999 to 2002, and the Vienna Research Center for Telecommunications, Vienna, Austria, until 2003. From 2003 to 2007, he was an Assistant Professor with the Mobile Communications Department, Institut Eurecom, Sophia Antipolis, France. In 2007, he was appointed as a Full Professor at Centrale-Supélec, Gif-sur-Yvette, France. From 2007 to 2014, he was the Director of the Alcatel-Lucent Chair on Flexible Radio. From 2014 to 2021, he was the Vice-President of the Huawei France Research Center, Boulogne-Billancourt, France, and jointly the Director of the Mathematical and Algorithmic Sciences Laboratory and the Lagrange Mathematical and Computing Research Center, Paris, France. Since 2021, he has been the Chief Research Officer with the Technology Innovation Institute, Abu Dhabi, United Arab Emirates. His research interests lie in fundamental mathematics, algorithms, statistics, information, and communication sciences research.

Dr. Debbah was a recipient of the ERC Grant MORE (Advanced Mathematical Tools for Complex Network Engineering) from 2012 to 2017. He received more than 20 best paper awards, including the Mario Boella Award in 2005, the IEEE Glavieux Prize Award in 2011, the Qualcomm Innovation Prize Award in 2012, the 2019 IEEE Radio Communications Committee Technical Recognition Award, and the 2020 SEE Blondel Medal.