This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

Fast MIMO Beamforming via Deep Reinforcement Learning for High Mobility mmWave Connectivity

Mahdi Fozi, Ahmad R. Sharafat, *Life Senior Member, IEEE*, and
Mehdi Bennis, *Fellow, IEEE*

***Abstract*—Future 5G/6G wireless networks will be increasingly using millimeter waves (mmWaves), where fast and efficient beamforming is vital for providing continuous service to highly mobile devices in the presence of interference and signal attenuation, manifested by blockage. In this paper, we propose a novel and efficient method for mmWave beamforming in massive multiple-input multiple-output (MIMO) systems to achieve the aforementioned goals with low complexity in such scenarios. In doing so, we utilize deep reinforcement learning (DRL) to maximize the network's energy efficiency subject to the quality of service (QoS) constraint for each user equipment (UE) and obtain its hybrid beamforming matrices. In doing so, we assume each UE is simultaneously associated with multiple access points (APs), i.e., simultaneous beamforming to/from multiple APs (coordinated multipoints) is needed for each UE. We also propose a low-complexity training algorithm, based on approximate message passing, which is well suited for the network edge. Besides, we develop a distributed scheme to reduce communications overhead via federated DRL. Extensive simulations show significant performance improvement over existing methods.**

***Index Terms*—High mobility, mmWave connectivity, hybrid beamforming, fast federated deep reinforcement learning, edge computing.**

I. Introduction

FUTURE NETWORKS are expected to provide new services such as virtual and augmented reality and vehicle-to-everything (V2X) communications [1] to highly mobile users, where millimeter waves (mmWaves) provide large amounts of bandwidth [2]. A key enabler for connecting a fast-moving user equipment (UE) to at least one access point (AP) at any instance is mitigating interference and steering beams (together called beamforming) in a timely and efficient manner with reasonable signaling overhead in the presence of channel variations and co-channel interference [2]–[4].

Fully digital beamforming is costly, power-hungry, and requires complex hardware [3]–[5], but hybrid beamforming can achieve comparable performance [6]–[9] with less cost and complexity. In hybrid beamforming, digital signal processing is employed in the baseband to eliminate/reduce interference, and discrete phase shifters are used in the RF to steer beams.

In hybrid structures, RF chains are typically group-connected to antennas via phase shifters. When the number of RF chains is the same as the number of antennas, energy consumption is high. When each RF chain is connected to all antennas (fully connected), interconnections are voluminous. In practice, RF chains are fewer than antennas and are connected to some (not all) antennas [9] to save energy and reduce interconnections. The existing beamforming schemes need channel state information (CSI) in a timely manner, obtained either by sparse channel estimation [10] or by exhaustive or hierarchical search [8], resulting in uncertain CSI or requiring excessive signaling, which are aggravated in high mobility cases, where channels and cell associations are fast changing.

We wish to develop a computationally efficient scheme for mmWave beamforming in the presence of channel variations and co-channel interference for fast-moving UEs. We consider multiuser and multicarrier networks, and optimize a given performance measure in partially-connected hybrid structures for mitigating interference and steering beams in a timely manner. In general, the problem is to minimize the distance between hybrid and fully digital beamforming [10] for each beam, which is known to be NP-hard [7]. In what follows, we briefly review prior works, and describe our contribution.

To reduce computations in optimization problems, various methods exist. The orthogonal matching pursuit is used in fully-connected structures, but with unsatisfactory results in partially-connected structures (PCSs) [10]. The alternating minimization method [7] is used for PCSs, but needs excessive computations in multiuser and multicarrier settings. Low complexity methods such as channel phase extraction [7] or convex relaxation [9] exist, but require excessive signaling to obtain uncertain CSI [11], [12]. To deal with uncertain CSI, QoS-aware schemes have been developed that either use CSI statistics (statistically robust) [13], [14], or consider an uncertainty region assumed to contain all instantaneous CSI (worst-case robust) [10], [11]. Nevertheless, all existing schemes fail to meet key features of high mobility communications as specified in the first release of 5G new radio (NR), e.g., in high-speed trains [15], [16].

Beamforming via deep supervised learning is a promising, scalable and statistically robust approach for high mobility cases [17]–[20]. In such schemes, RF signature of the environment and locations of users/APs are obtained via pilot signals [17], and contextual side-information such as user trajectory [18], past beamforming [17], [20]–[22], situational awareness [23] and traffic flow [24] are used in the training phase. The trained model is then used for online beamforming to connect each fast-moving UE with at least one AP. Various deep learning paradigms, such as the generative adversarial estimation

Manuscript received June 10, 2021, revised September 10, 2021.

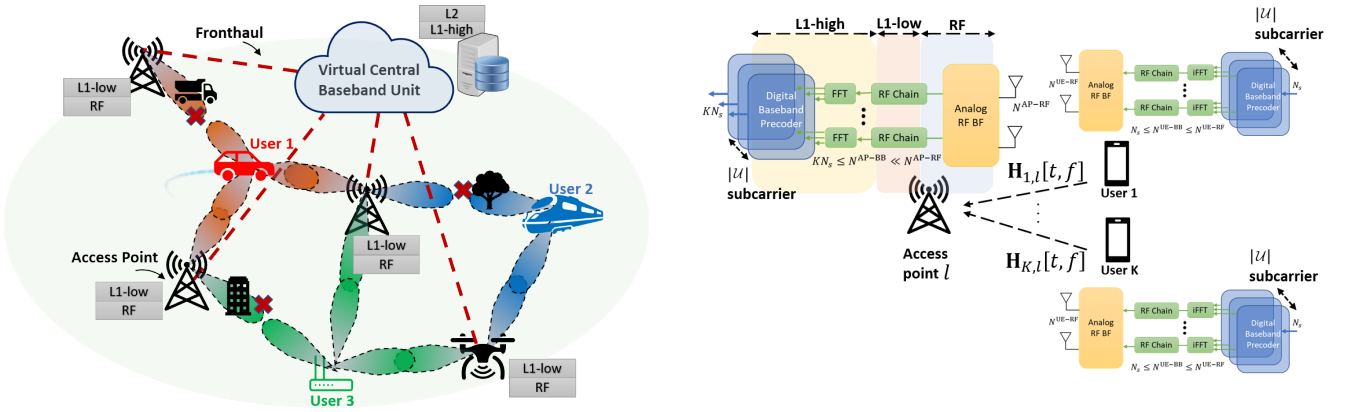
M. Fozi and A. R. Sharafat are with Tarbiat Modares University, Tehran, Iran. M. Bennis is with the University of Oulu and Academy of Finland.

Corresponding author is Ahmad R. Sharafat (e-mail: sharafat@ieee.org).

This work is supported in part by Academy of Finland 6G Flagship (grant no. 318927), project SMARTER, projects EU-ICT IntellIoT and EUCHISTERA LearningEdge, and CONNECT, Infotech-NOOR, and NEGEIN.

Digital Object Identifier 10.1109/JSAC.2021.xxxxxx

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications



(a) (b)

Figure 1. (a) Network layout in which each fast-moving UE is simultaneously served by multiple APs, i.e., coordinated multipoint, depending on the service and fronthaul load. Typical cell radius for urban (rural) deployments are 150 m (580 m), which means that each user stays only 10.8 seconds (8.35 seconds) in the footprint of each AP for speeds of 100km/h (500km/h) [15]. (b) Block diagram of the system.

of channel covariance [20], long short-term memory in single-user scenarios [21] and deep convolutional neural networks in the downlink of multi-user settings [22] have been proposed.

In general, performance of *supervised* deep learning algorithms is promising [25], but require extensive labeled datasets for training, and are sensitive to unpredictable variations in mmWave channels manifested by prevalent blockage [26]. To alleviate this, in [27], deep reinforcement learning (DRL) is used for hybrid beamforming in point-to-point communications. DRL has its costs as well: Its convergence is slow and needs excessive computations, usually provided via cloud computing with high latency and excessive signaling [17]. To manage slow convergence, we present novel DRL-based schemes with reduced convergence time. Besides, stringent time constraints in fast-moving UEs can be met by utilizing edge computing (with significantly less signaling and reduced mobility interruption time) instead of cloud computing [28], but the challenge is scarceness of computing power at the edge.

Efficient, fast, and low-overhead beamforming in the presence of unpredictable line-of-sight (LoS) blockage and channel uncertainties in space, frequency, and time in mmWave bands is needed to connect a fast-moving UE with at least one AP with strict limits on beam steering latency, as shown in Fig. 1(a). In this paper, we present a novel DRL-based approach with low-overhead training for fast hybrid beamforming in urban and rural deployments. Specifically, we present a centralized and a distributed processing/training scheme for DRL, both of which can achieve the above objectives. Our schemes avoid the problem of frequent handovers of fast-moving UEs that are in cell-boundaries and/or are not in LoS (NLoS).

In our centralized processing/training scheme, the weights are learned by the DRL agent by alternating between training and online beamforming. In general, centralized schemes suffer from high communications overhead and require phase synchronization among APs. To reduce communications overhead, we develop a distributed processing/training scheme by utilizing federated learning, which does not require phase synchronization as well, but the challenge is utilizing locally processed data to obtain the shared optimum weight values.

Our contributions can be summarized as:

- We show how DRL can be used for fast beamforming in mmWave massive multiple-input multiple-output (MIMO) channels in high mobility communications. In doing so, we develop an efficient, practical, and convergent centralized processing/training algorithm whose performance is stable in the presence of significant variations in the UE's velocity and uncertainties in typical values of CSI.
- We also develop a distributed processing/training scheme whose communications overhead is significantly less, i.e., is fronthaul-load scalable, and does not need phase synchronization among APs.
- We apply our schemes in two important use cases, namely vehicle to infrastructure (V2I) and high-speed train (HST) communications in a train to infrastructure (T2I) scenario in ultra wideband mmWave bands with spatial non-stationarity in massive MIMO channels, and benchmark our schemes against other existing schemes that do not require perfect CSI. We also show that our approach has important practical benefits.

The following notations are used in this paper: $\mathbf{A}$, $\mathbf{a}$, a, $\mathcal{A}$, and A denote a matrix, a vector, a scalar, a set, and a function, respectively; $a \propto b$ denotes proportionality; $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$; $[\mathbf{A}]_{i,j}$ denotes the (i,j)-th entry of matrix $\mathbf{A}$; $\mathbf{F}$ and $\mathbf{W}$ are the uplink hybrid precoder and combiner matrices, respectively; $\mathbf{H}$ denotes the channel between an AP and a mobile unit; $\mathcal{N}_{\mathbb{C}}(\mathbf{m}, \mathbf{C})$ denotes a complex normal distribution with mean $\mathbf{m}$ and covariance $\mathbf{C}$; $\mathbb{E}_x(\cdot)$ denotes the ensemble average with respect to x; $(\cdot)^{\mathsf{H}}$ is the Hermitian operator; $\iota = \sqrt{-1}$; and $\mathbb{P}(x)$ is the probability of event x.

This paper is organized as follows. The system model and channel model are described in Section II. In Sections III and IV, the problem is formulated and our proposed schemes are described, respectively. In Section V, performance of our schemes is numerically studied and compared with those of other existing methods, followed by conclusion in Section VI.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

II. System Model and Channel Model

A. System Model

Consider L APs in the network, numbered from 1 to L, each equipped with $N^{\text{AP-RF}}$ antennas and $N^{\text{AP-BB}} \ll N^{\text{AP-RF}}$ RF chains. There are K UEs, each with $N^{\text{UE-RF}}$ antennas and $N^{\text{UE-BB}} \leq N^{\text{UE-RF}}$ RF-chains. Each UE k is simultaneously served by multiple APs in $\mathcal{M}_k \subset \{1,\ldots,L\}$ over shared bandwidth W, as shown in Fig. 1(b). The channel between UE k and AP l is denoted by $\mathbf{H}_{k,l} \in \mathbb{C}^{N^{\text{AP-RF}} \times N^{\text{UE-RF}}}$. Assume channel reciprocity in time division duplexing (TDD). Hence, the estimate of UL channel at each AP can be readily used in the downlink (DL) after compensating for any mismatches between the transmitter-receiver pair. Each UE k sends UL pilot signals, which enable AP l to locally estimate its channel to that UE, denoted by $\hat{\mathbf{H}}_{kl}$.

Let $\mathbf{s}_k \in \mathbb{C}^{N_s}$ with $N_s \leq N^{\text{UE-BB}}$ denote the normalized (unit-power) signal that UE k wants to transmit, where N_s is the number of independent transmit streams by that UE. In orthogonal frequency division multiplexing (OFDM), each UE k modulates its data stream $\mathbf{s}_k$ by taking a $|\mathcal{U}|$-point IFFT and adding a cyclic prefix of length D to obtain $\mathbf{s}_k[t,f]$ for subcarrier $f \in \mathcal{U}$ at discrete instance t. The total transmit power of UE k in shared channels to all APs in $\mathcal{M}_k$ is $p_k[t,f] \geq 0$, obtained via the water-filling algorithm. From an interference perspective, this is a worst-case scenario. Assuming perfect frequency and carrier offset synchronization, the received signal at AP l denoted by $\mathbf{y}_l^{\text{UL}} \in \mathbb{C}^{N^{\text{AP-RF}}}$ is [29]

$$\mathbf{y}_l^{\text{UL}}[t,f] = \sum_{k=1}^{K} \mathbf{H}_{k,l}[t,f]\mathbf{F}_k[t,f]\sqrt{p_k[t,f]}\mathbf{s}_k[t,f] + \mathbf{n}_l[t,f], \quad (1)$$

where $\mathbf{F}_k[t,f] = \mathbf{F}_k^{\text{RF}}[t]\mathbf{F}_k^{\text{BB}}[t,f]$ is the radio front end for UE k, and $\mathbf{F}_k^{\text{RF}}$ and $\mathbf{F}_k^{\text{BB}}$ are the RF and baseband precoding matrices of UE k, respectively, $\mathbf{n}_l \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{N^{\text{AP-RF}}}, \sigma^2\mathbf{I}_{N^{\text{AP-RF}}})$ is the complex-valued independent additive white Gaussian noise, and $\mathbf{I}_N$ is the $N \times N$ identity matrix. Note that the UL RF beamformer for each UE is common to all subcarriers.

APs in $\mathcal{M}_k$ use their received signals $\{\mathbf{y}_l^{\text{UL}} : l \in \mathcal{M}_k\}$ to jointly detect the signal received from UE k. As shown in Fig. 1(b), each AP $l \in \mathcal{M}_k$ selects the baseband and RF beamformer matrices for UE k, denoted by $\mathbf{W}_{l,k}^{\text{BB}}$ and $\mathbf{W}_l^{\text{RF}}$, respectively, and computes

$$\mathbf{r}_{l,k}[t,f] = \mathbf{W}_{l,k}[t,f]\mathbf{y}_l^{\text{UL}}[t,f], \quad (2)$$

where $\mathbf{W}_{l,k}[t,f] = \mathbf{W}_{l,k}^{\text{BB}}[t,f]\mathbf{W}_l^{\text{rf}}[t]$. Note that the analog receive beamformer in AP $l \in \mathcal{M}_k$, i.e., $\mathbf{W}_l^{\text{RF}}[t]$, is common to each UE k and all subcarriers. The values of $\mathbf{r}_{l,k}$ are sent to the virtual central baseband unit which takes L $|\mathcal{U}|$-point FFTs and combines them to get

$$\mathbf{r}_k[t,f] = \mathbf{G}_{k,k}[t,f]\sqrt{p_k}[t,f]\mathbf{s}_k[t,f] + \sum_{j\neq k} \mathbf{G}_{k,j}[t,f]\sqrt{p_j}[t,f]\mathbf{s}_j[t,f] + \mathbf{m}_k[t,f], \quad (3)$$

where $\mathbf{r}_k = \sum_{l\in\mathcal{M}_k}\mathbf{r}_{l,k}$, $\mathbf{G}_{k,j} = \sum_{l\in\mathcal{M}_k}\mathbf{W}_{l,k}\mathbf{H}_{j,l}\mathbf{F}_j$ and $\mathbf{m}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{N^{\text{AP-BB}}}, \sigma^2\mathbf{C}_m)$ where $\mathbf{C}_m = \sum_{l\in\mathcal{M}_k}\mathbf{W}_{l,k}\mathbf{W}_{l,k}^{\text{H}}$ is the post-processed colored-noise. We rewrite (3) as

$$\mathbf{R}[t,f] = \mathbf{S}[t,f]\mathbf{G}[t,f] + \mathbf{M}[t,f], \quad (4)$$

where $[\mathbf{G}]_{k,j} = \sqrt{p_j}\mathbf{G}_{k,j}$.

From [29, Theorem 4.1], a tight lower bound on the achievable spectral efficiency (SE) for UE k in nats/s/Hz is

$$\text{SE}_k^{\text{UL}} \gtrapprox \frac{1}{|\mathcal{U}|}\sum_{f\in\mathcal{U}} \log\det\left(\mathbf{I} + \boldsymbol{\mathcal{G}}_k^{\text{UL}}[f]\right), \quad (5)$$

where $\boldsymbol{\mathcal{G}}_k^{\text{UL}}[f]$ is given by (6).

This lower bound is achieved by utilizing minimum mean squared error with successive interference cancellation (MMSE-SIC). Note that $\|\boldsymbol{\mathcal{G}}_k^{\text{UL}}\|_2^2$ is the UL effective signal-to-interference-plus-noise ratio (SINR). When UEs have the same priority, UL SE in nats/s/Hz is [30]

$$\text{SE}^{\text{UL}} = \log\det\left(\mathbf{I} + \mathbf{C}_{\mathbf{M}}^{-\text{H}/2}\mathbf{G}^{\text{H}}\mathbf{G}\mathbf{C}_{\mathbf{M}}^{-1/2}\right) \gtrapprox \sum_{k=1}^{K}\text{SE}_k^{\text{UL}}. \quad (7)$$

Similar results hold for DL SE [31]. The total UL consumed power P^{UL} is the sum of all UEs' transmit power and the static hardware power consumed in all APs and UEs, i.e.,

$$P^{\text{UL}} = K \times P_{\text{UE-static}} + L \times P_{\text{AP-static}} + \sum_{f\in\mathcal{U}}\sum_{k=1}^{K}\eta_k^{-1}p_k[f], \quad (8)$$

where η_k is the efficiency of power amplifier in mobile unit k, and $P_{\text{UE-static}}$ and $P_{\text{AP-static}}$ are the static hardware power consumption by one mobile unit and one AP, respectively. In the above, we assume that transmit amplifiers operate in their linear region and static hardware power consumption is the same irrespective of data rates.

Energy efficiency (EE) is defined as the ratio of the system's spectral efficiency in nats/s/Hz to the total power consumption in Joule/s (Watt) for a given bandwidth of W, i.e.,

$$\text{EE}^{\text{UL}} = W\frac{\text{SE}^{\text{UL}}}{P^{\text{UL}}} \quad \text{nats/Joule}. \quad (9)$$

B. Channel Model

Consider one transmitter (UE), one receiver (AP), and a cluster of scatterers between UE and AP. We adopt a 3D time-varying wideband geometry-based stochastic channel model between UE k and AP l where the $N^{\text{AP-RF}} \times N^{\text{UE-RF}}$ frequency-domain baseband channel transfer function (CTF) is [32], [33]

$$[\mathbf{H}_{l,k}(t,f)]_{i,j} = \sqrt{\frac{K_{i,j}(t)}{K_{i,j}(t)+1}}[\mathbf{H}_{l,k}^{\text{LoS}}(t,f)]_{i,j} + \sqrt{\frac{1}{K_{i,j}(t)+1}}\sum_{n=1}^{N(t)}[\mathbf{H}_{l,k,n}^{\text{NLoS}}(t,f)]_{i,j}, \quad (10)$$

where $K_{i,j}(t)$ is the K-Ricean factor, $[\mathbf{H}_{l,k}^{\text{LoS}}(t,f)]_{i,j}$ is the LoS component given by (11), and $[\mathbf{H}_{l,k,n}^{\text{NLoS}}(t,f)]_{i,j}$, as given by (12), is the narrow-band process associated with all M_n irresolvable sub-paths in each cluster that have the same delay $\tau_{i,j}^{n,m}(t)$ and mean gain $\alpha_{n,m}(t)$ (including path loss and shadowing). The channel is modeled by a two-state Markov process whose state depends on the LoS blockage. The azimuth and elevation angles of arrival (AoAs) of sub-path m in cluster n at receive antenna i are $\phi_{i,m,n}^{\text{Rx}}(t)$ and $\theta_{i,m,n}^{\text{Rx}}(t)$, respectively. Similarly, the azimuth and elevation angles of departure (AoDs) of sub-path m in cluster n from transmit antenna j are $\phi_{j,m,n}^{\text{Tx}}(t)$ and $\theta_{j,m,n}^{\text{Tx}}(t)$, respectively. The azimuth (elevation) AoAs/AoDs are assumed to have wrapped Gaussian (truncated Laplacian) distribution whose

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

$$\boldsymbol{\mathcal{G}}_k^{\mathrm{UL}}[f] = \left(\sum_{j\neq k} p_j \mathbf{G}_{k,j}^{\mathsf{H}} \mathbf{G}_{k,j} + \sigma^2 \mathbf{C}_m\right)^{-\mathsf{H}/2} p_k \mathbf{G}_{k,k}^{\mathsf{H}} \mathbf{G}_{k,k} \left(\sum_{j\neq k} p_j \mathbf{G}_{k,j}^{\mathsf{H}} \mathbf{G}_{k,j} + \sigma^2 \mathbf{C}_m\right)^{-1/2}. \tag{6}$$

$$\begin{aligned}[\mathbf{H}_{l,k}^{\mathrm{LoS}}(t,f)]_{i,j} = &\begin{bmatrix} \mathsf{F}_{i,V}^{\mathrm{Tx}}\left(\phi_{i,\mathrm{LoS}}^{\mathrm{Tx}}(t), \theta_{i,\mathrm{LoS}}^{\mathrm{Tx}}(t)\right) \\ \mathsf{F}_{i,H}^{\mathrm{Tx}}\left(\phi_{i,\mathrm{LoS}}^{\mathrm{Tx}}(t), \theta_{i,\mathrm{LoS}}^{\mathrm{Tx}}(t)\right) \end{bmatrix}^H \begin{bmatrix} e^{\iota\gamma_{V,V}^{\mathrm{LoS}}} & 0 \\ 0 & e^{\iota\gamma_{H,H}^{\mathrm{LoS}}} \end{bmatrix} \\ &\times \begin{bmatrix} \mathsf{F}_{j,V}^{\mathrm{Rx}}\left(\phi_{j,\mathrm{LoS}}^{\mathrm{Rx}}(t), \theta_{j,\mathrm{LoS}}^{\mathrm{Rx}}(t)\right) \\ \mathsf{F}_{j,H}^{\mathrm{Rx}}\left(\phi_{j,\mathrm{LoS}}^{\mathrm{Rx}}(t), \theta_{j,\mathrm{LoS}}^{\mathrm{Rx}}(t)\right) \end{bmatrix} \times e^{-\iota 2\pi \nu_{i,j}^{\mathrm{LoS}}(t) t} e^{-\iota 2\pi f \tau_{i,j}^{\mathrm{LoS}}(t)} e^{-\iota \frac{2\pi f}{c} D_{i,j}^{\mathrm{LoS}}(t)}\end{aligned} \tag{11}$$

$$\begin{aligned}[\mathbf{H}_{l,k,n}^{\mathrm{NLoS}}(t,f)]_{i,j} = &\sum_{m=1}^{M_n} \left(\frac{f}{f_c}\right)^{\beta_{n,m}} \sqrt{\frac{\alpha_{n,m}(t)}{M_n}} \begin{bmatrix} \mathsf{F}_{i,V}^{\mathrm{Tx}}\left(\phi_{i,m,n}^{\mathrm{Tx}}(t), \theta_{i,m,n}^{\mathrm{Tx}}(t)\right) \\ \mathsf{F}_{i,H}^{\mathrm{Tx}}\left(\phi_{i,m,n}^{\mathrm{Tx}}(t), \theta_{i,m,n}^{\mathrm{Tx}}(t)\right) \end{bmatrix}^H \\ &\times \begin{bmatrix} \frac{1}{\sqrt{\kappa_{n,m}}} e^{\iota\gamma_{V,V}^{n,m}} & e^{\iota\gamma_{V,H}^{n,m}} \\ e^{\iota\gamma_{H,V}^{n,m}} & \frac{1}{\sqrt{\kappa_{n,m}}} e^{\iota\gamma_{H,H}^{n,m}} \end{bmatrix} \begin{bmatrix} \mathsf{F}_{j,V}^{\mathrm{Rx}}\left(\phi_{j,m,n}^{\mathrm{Rx}}(t), \theta_{q,m,n}^{\mathrm{Rx}}(t)\right) \\ \mathsf{F}_{j,H}^{\mathrm{Rx}}\left(\phi_{j,m,n}^{\mathrm{Rx}}(t), \theta_{j,m,n}^{\mathrm{Rx}}(t)\right) \end{bmatrix} \\ &\times e^{-\iota 2\pi \nu_{i,j}^{n,m}(t) t} e^{-\iota 2\pi f \tau_{i,j}^{n,m}(t)} e^{-\iota \frac{2\pi f}{c}} D_{i,j}^{n,m}(t)\end{aligned} \tag{12}$$

parameter values are known for each scenario [34]. The functions $\mathsf{F}_V^{\mathrm{Tx}}(\mathsf{F}_H^{\mathrm{Tx}})$ and $\mathsf{F}_V^{\mathrm{Rx}}(\mathsf{F}_H^{\mathrm{Rx}})$ denote the antenna patterns of vertical (horizontal) polarization of transmit and receive arrays, respectively. The transmit and receive array response vectors are $\mathbf{a}_{\mathrm{Tx}}(\cdot,\cdot)$ and $\mathbf{a}_{\mathrm{Rx}}(\cdot,\cdot)$, respectively.

When a uniform linear array (ULA) with I antenna elements and d spacing is employed at the receiver broadside, we have $[\mathbf{a}_{\mathrm{Rx}}]_i = e^{-\iota(i-1)d\kappa \sin\phi_{m,n}^{\mathrm{Rx}}(t)}$ for $i = 1, \ldots, I$, where $\kappa = \frac{2\pi}{\lambda_c}$ is the wave number and λ_c is the carrier wavelength. Also, $\nu_{m,n}^{k,l}(t) = \frac{\kappa}{2\pi}\mathbb{R}(\mathbf{v}_{k,l}^{\mathsf{H}} \mathbf{e}_{m,n})$ is the Doppler frequency where $\mathbf{v}_{k,l}$ is the relative velocity vector of UE k and AP l, $\mathbf{e}_{m,n}$ is the viewing direction vector of sub-path m in cluster n toward UE, $\mathbb{R}(\cdot)$ returns the real part of a complex scalar, $\beta_{n,m}$ is the frequency-dependent factor, and $\kappa_{n,m}$ denotes the cross polarization power ratio. Finally, $\gamma_{V,V}^{m,n}$ is a random phase uniformly distributed in $[-\pi, \pi)$ associated with scatterer m in cluster n in the vertical-vertical polarization, and $\gamma_{V,H}^{m,n}, \gamma_{H,V}^{m,n}$, and $\gamma_{H,H}^{m,n}$ are similarly defined. The last term denotes the group delay of each path.

In high-mobility communications with distributed large scale massive MIMO, channels are assumed to be wideband and non-stationary in space and time [32]. These features are considered in the channel model in (11). Due to movement of each UE k, AP l, and the cluster of scatterers, parameter values of channel model are time-varying, but assumed to be stationary in short intervals during which fading statistics remain invariant. We use the procedure described in [33] to generate channel parameters for each stationary interval.

In mmWaves, channels are typically sparse in the angular domain resulting in few, say 3-5 paths [2] which may experience blockage with the following probability

$$\begin{aligned}\mathbb{P}(\text{Blockage}) &= \mathbb{P}(\text{NLoS}) \times \mathbb{P}(N(t) = 0 | \text{NLoS}) \\ &= \frac{1}{1 + \mathbb{E}(K_{i,j})} \times (1 - \frac{\lambda_b}{\lambda_d}),\end{aligned} \tag{13}$$

where λ_b and λ_d are the birth rate and death rate of the birth-death process associated with $N(t)$, respectively [33].

We focus on cases where fading coefficients vary quickly; i.e., timely and accurate estimation of coefficients is not feasible. We model the channel as a stochastic process assumed to be stationary over time T, where

$$T = \min\{T_b, T_c\}, \tag{14}$$

in which T_b is the beam coherence time and T_c is the channel coherence time, both of which depend on UEs' mobility and channel multipath parameters [35]. In practice, $T_b \gg T_c$, and $T^{\mathrm{NLoS}} \propto \frac{1}{\Theta^\alpha \times f_D}$ and $T^{\mathrm{LoS}} \propto \frac{1}{\sin\phi \times \Theta \times f_D}$, for NLoS and LoS, respectively, where f_D is the maximum Doppler frequency, Θ is the mean beamwidth, $\alpha \in \{1, 2\}$ is a scenario-dependent parameter, and ϕ is the direction toward the transmitter. Note that T is higher in LoS due to the beamforming gain, e.g., the coherence time is around 23 ms for a vehicle moving at 48 km/h [36], and is 9 ms for a HST moving at 324 km/h [37]. We use these values in Section V.

III. Problem Statement

Without loss of generality, we focus on the uplink (UL). Ideally, each UE k simultaneously steers its beams towards APs in $\mathcal{M}_k$ with optimal transmit power to maximize spectral/energy efficiency. At the same time, each AP simultaneously steers its receive beams towards its corresponding UEs. In this context, fast beamforming becomes vital. We assume that exact CSI is unavailable since obtaining exact CSI in a timely manner is nontrivial and costly. We also assume hybrid beamforming, where RF beamformers are shared among multiple UEs and subcarriers, and UEs may have different QoS requirements (SE or EE). In noise-limited paradigms, SE and EE are aligned, which is not the case in interference-limited paradigms. The way in which resources are allocated and interference is mitigated significantly affect beamforming.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

A. Problem Formulation

Considering the beam steering latency, power consumption by UE and its throughput, we wish to maximize the QoS-aware UL EE, i.e.,

$$\underset{\substack{\tau_{\text{train}},\\ \mathbf{F}_k[f]\forall f\forall k,\\ \mathbf{W}_{l,k}[f]\forall f\forall k\forall l}}{\text{maximize}} \quad \text{EE}^{\text{UL}} \tag{15a}$$

$$\text{subject to} \begin{cases} \left(1-\frac{\tau_{\text{train}}}{T}\right)\text{SE}_k^{\text{UL}} \geq R_k \geq 0 & \forall k, \\ \mathbf{W}_l^{\text{RF}} \in \mathcal{W} & \forall l, \\ \mathbf{F}_k^{\text{RF}} \in \mathcal{F} & \forall k, \\ \|\mathbf{W}_{l,k}[f]\|_F^2 = N^{\text{AP-BB}} & \forall f, \forall k, \forall l, \\ \|\mathbf{F}_k[f]\|_F^2 = N^{\text{UE-BB}} & \forall f, \forall k, \end{cases} \tag{15b}$$

where $\tau_{\text{train}} \leq T$ is the training length, R_k is the minimum required data rate per bandwidth for UE k, and $\mathcal{W}$ and $\mathcal{F}$ are the sets of RF beamforming matrices and RF precoding matrices of APs and UEs, respectively.

The first constraint in (15b) makes the system QoS-aware by requiring each UE's data rate per bandwidth be higher than its minimum required value. The second constraint in (15b) limits the per-subcarrier transmit power. Also, the Frobenius norm of precoding and beamforming matrices in the last two constraints in (15b), limit the consumed power. The same power budget is assumed for both training and operation. The above formulation considers simultaneous associations to multiple APs for each user. Solving (15) in wideband communications requires many calculations because of coupling between subcarriers. To reduce calculations, one may decouple subcarriers by utilizing block diagonalization precoders, but as we will show in Section V, this reduces the achievable rate.

The virtual central baseband unit solves (15) to find the optimal beamforming matrices and transmit power vectors in each short interval T during which while the channel is assumed to be stationary but unknown, the beams should be steered. Since (15) is a non-convex problem and has mixed discrete/continuous variables, an efficient and at least asymptotically optimal method is needed to decouple the transmitter design from the receiver design and obtain their respective matrices [7]. This significantly reduces the problem size. Note that DRL interacts well with unpredictable and unknown environments, e.g., high mobility mmWave communications, by alternating between exploration (training) and exploitation (operation) to maximize a cumulative reward while solving consecutive instances of (15). However, to use DRL for real-time mmWave beamforming for highly mobile users, a low-complexity training algorithm is needed.

B. Preliminaries

The partially observed Markov decision process (POMDP) is a mathematical framework that models interactions of an agent with an unknown time-varying environment when the agent has limited observations. POMDP is a sextuple $(\mathcal{S}, \mathcal{A}, \mathsf{P}, \mathsf{R}, \Omega, \mathsf{O})$, where $\mathcal{S}$ is the set of environment states, $\mathcal{A}$ is the set of agent's actions, $\mathsf{P}: \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ is the state transition function, i.e., $\mathbb{P}(s_{t+1}|s_t,a_t)$, $\mathsf{R}: \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the reward function whose expectation is $\mathbb{E}_R\{\mathsf{R}(s_t,a_t)|s_t,a_t\}$, Ω is the set of limited observations, and $\mathsf{O}: \mathcal{S}\times\mathcal{A}\to\Delta(\Omega)$ is the observation function, i.e., $\mathbb{P}(o|s_{t+1},a_t), \forall o\in\Omega$. The Markov process entails that future reward values depend only on past history of states and actions, i.e., $\mathbb{P}(r_t, s_{t+1}|s_0,a_0,r_0,\cdots,s_t,a_t) = \mathbb{P}(r_t,s_{t+1}|s_t,a_t)$, which is often the case in practice.

In POMDP, the agent adopts an optimal nonstationary policy denoted by $\pi^*: \Omega\to\Delta(\mathcal{A})$ to maximize the expected reward. Typically, for a POMDP, a belief state is defined as

$$b_\pi(s) = \mathbb{P}(s_t|I_t^{\text{C}}), \tag{16}$$

where I_t^{C} is the complete sufficient information state at time t. The value of a belief b under policy π, denoted by $V_\pi(b)$, is the expected return when believing b and following π thereafter.

It is well-known that the optimal solution $V_\pi^*(b)$ satisfies the following Bellman optimality equation [38]

$$V_\pi^*(b) = \max_{a\in\mathcal{A}}\left(\mathsf{R}(b,a) + \gamma\sum_{b'}\mathbb{P}(b'|b,a)V^*(b')\right), \tag{17}$$

where $\gamma\in(0,1]$ is the weight factor for the sum of future rewards, $\mathsf{R}(b,a) = \sum_s b(s)\mathsf{R}(s,a)$ and $\mathbb{P}(b'|a,b) = \sum_{o',s',s}\mathbb{P}(b'|a,b,o')\mathbb{P}(o'|s',a)\mathbb{P}(s'|s,a)b(s)$. The belief updating in (17) can be computed only for discrete low-dimensional $\mathcal{S}$ and linear-Gaussian dynamics. The model-free reinforcement learning (RL) can overcome the above challenge, where the agent explores the state space to tune its action on a trial-and-error basis [38].

IV. Proposed Method

In this section, we develop a framework to use POMDP-based DRL for solving (15). We also develop a low complexity training algorithm for our proposed scheme. An important issue for solving (15) is whether CSI is needed. Theorem IV.1 below shows that when SINR is high, CSI may not be needed (which is desirable) and noncoherent multiuser communications can be considered. The need for less computations and resources when CSI is not needed is in fact the motivation behind our proposed framework.

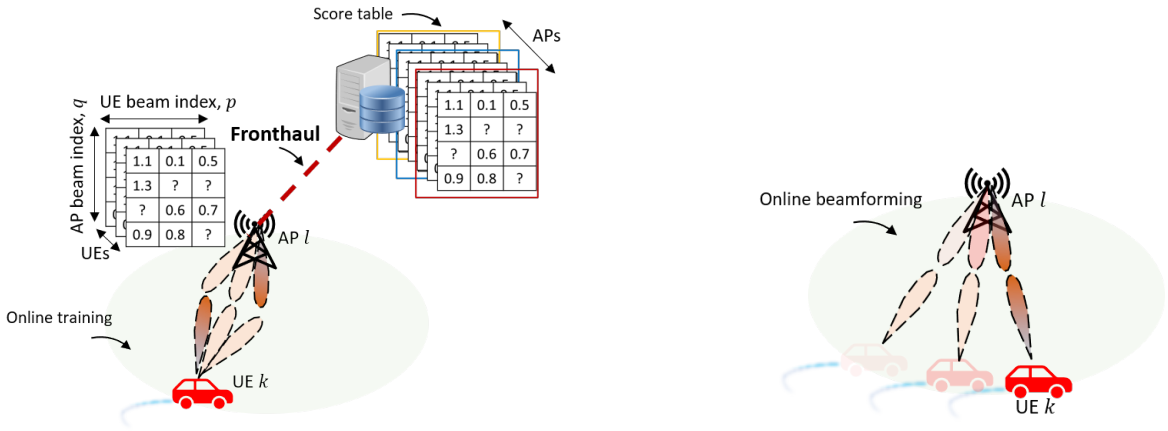
Theorem IV.1. *Problem* (15) *asymptotically has a solution iff* $(R_1,\ldots,R_K)\in\mathcal{C}(P_{k,\max}\forall k)$ *where* $\mathcal{C}(P_{k,\max}\forall k)$ *is the system's polymatroid noncoherent capacity region and* $P_{k,\max}$ *is the maximum transmit power of UE* k.

Proof. Given a common diversity denoted by d for users, we find a set of K-tuple achievable multiplexing gains $(r_1,\ldots,r_K)$, denoted by $\mathcal{R}(d)$, which is a polymatroid whose rank function is

$$f(S) = \begin{cases} |S| r_{m,n}^*(d), & \text{if } 0\leq|S|\leq l-1 \\ r_{|S|m,n}^*(d), & \text{if } l\leq|S|\leq K, \end{cases}$$

where $d\in[d_{l-1},d_l]$, $d_l = d_{m,n}^*(\frac{n}{K+1})$ and $d_{m,n}^*(n)$ is the single-user noncoherent rate-diversity tradeoff given by sphere packing in the Grassmann manifold, i.e., $\mathcal{C} = \mathsf{DoF}\log_2\mathsf{SINR} + c(N^{\text{UE-BB}}, \frac{L}{K}N^{\text{AP-BB}}, T_q) + o(1)$ [39], where $\mathsf{DoF} = M^*(1-\frac{M^*}{T_q})$, $M^* = \min\{\frac{L}{K}N^{\text{AP-BB}}, N^{\text{UE-BB}}, \lfloor\frac{T_q}{2}\rfloor\}$, $o(1)$ is a vanishing term as $\mathsf{SINR}\to\infty$, and T_q is the quantized coherence time T. Hence, when SINR is high, there is no need to have CSI. With beamforming, channel gain (and SINR) is the highest in the desired direction. At high SINR, using only M^* of the $N^{\text{AP-BB}}$ available RF chains is optimal, i.e., using more transmit antennas than receive RF chains does not yield

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications



(a) (b)

Figure 2. All connections (beam pairs) between AP l and UE k are (a) trained via orthonormal UL pilots among which the connection(s) giving the highest score (i.e., total achievable rate (received energy) across all subcarriers) are chosen for (b) online beamforming. In high mobility communications, obtaining all scores is not practical, and one should consider only the observed scores (not the complete set) and predict the remaining ones.

any capacity increase. A capacity achieving scheme is to use Grassmannian signaling [40], [41]. □

Similar to [10], in (15), we separately consider transmit precoding and receive beamforming for each UE k. To obtain the transmit precoding matrix, we use primal-dual interior point methods to solve

$$\underset{\mathbf{F}_k:\mathbf{F}_k^{\text{RF}}\in\mathcal{F}}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt},k}-\mathbf{F}_k\|_F+\lambda_k(R_k-\text{SE}_k^{\text{UL}}), \qquad (18)$$

where $\mathbf{F}_{\text{opt},k}$ is the transmit precoding matrix for the fully-digital design, and $\lambda_k(\cdot)$ is the logarithmic barrier. We then use the precoding matrices $\mathbf{F}_k^{(t)}$ to obtain the receive beamforming matrix $\mathbf{W}_{l,k}^{(t)}$ $\forall l$ $\forall k$ by minimizing mean squared error (MSE). We cycle through the above until either convergence or termination. To solve (18), we separately consider RF beamforming and baseband beamforming.

A. RF beamforming via POMDP-Based DRL

A connection between UE k and AP l involves two beams (one for the uplink and one for the downlink), namely, $\mathbf{F}_k^{\text{RF}}\in\mathcal{F}$ and $\mathbf{W}_l^{\text{RF}}\in\mathcal{W}$. Hence, the connection space is $\mathcal{F}^K\times\mathcal{W}^L$ from which the connections with maximum received energy (achievable rate) are chosen after training is completed. When CSI is available at no cost, there is no need for training, and $\tau_{\text{train}}^{\text{optimal}}=0$. In practice, however, obtaining CSI is costly and nontrivial. In this case, as shown in Fig. 2, each UE k repeatedly transmits $|\mathcal{W}|$ known (orthonormal) UL pilot sequences $\mathbf{S}_k^{\text{Pilot}}[f]\in\mathbb{C}^{N^{\text{UE-BB}}\times T_P}$ for each beam in $\mathcal{F}$, during which each AP l cycles through its RF beamforming matrices and combines every received pilot sequence with a different RF beamforming matrix.

Each AP (UE) has $N^{\text{AP-RF}}(N^{\text{UE-RF}})$ beams out of which at most $N^{\text{AP-BB}}(N^{\text{UE-BB}})$ is active. After multiplying the received RF signal by $(\mathbf{S}_k^{\text{Pilot}}[f])^{\text{H}}$, the baseband received signal that corresponds to the q-th RF beamforming matrix of AP l and the p-th RF beamforming matrix of UE k is

$$\mathbf{y}_{l,q,k,p}^{\text{Pilot}}[f]=\mathbf{W}_{l,q}^{\text{RF}}\mathbf{H}_{k,l}[f]\mathbf{F}_{k,p}^{\text{RF}}+\text{noise}. \qquad (19)$$

Next, AP l calculates its total achievable rate (*score*) for UE k as

$$\sum_{f\in\mathcal{U}}\log\det(\mathbf{I}+\mathbf{C}_m^{-1}\mathbf{F}_{k,p}^{\text{RF}\,\text{H}}\mathbf{H}_{k,l}[f]^{\text{H}}\mathbf{W}_{l,q}^{\text{RF}\,\text{H}}\mathbf{W}_{l,q}^{\text{RF}}\mathbf{H}_{k,l}[f]\mathbf{F}_{k,p}^{\text{RF}}).$$

In doing so, in (19), AP l estimates channels $\mathbf{H}_{k,l}[f],\forall k\forall f$ by assuming known $\mathbf{W}_{l,q}^{\text{RF}}$ and $\mathbf{F}_{k,p}^{\text{RF}}$, and observing $\mathbf{y}_{l,q,k,p}^{\text{Pilot}}[f]$. When noise is low, it may use suboptimal RF energy estimator. A set of RF beams, say $\{\mathbf{W}_{l,q}^{\text{RF}}\forall l,\mathbf{F}_{k,p}^{\text{RF}}\forall k\}$, is QoS-aware when the sum of their associated scores from all APs is no less than R_k, for all k (the first constraint in (15b)). All scores are then sent to the virtual central baseband unit for solving (18) by searching through all QoS-aware sets of RF beams that maximizes (15a). The search space exponentially grows with input size, i.e.,

$$\mathcal{O}\left(\left(\frac{(N^{\text{UE-RF}})^{N^{\text{UE-BB}}}}{N^{\text{UE-BB}}!}\right)^K\times\left(\frac{(N^{\text{AP-RF}})^{N^{\text{AP-BB}}}}{N^{\text{AP-BB}}!}\right)^L\right).$$

In high mobility communications, obtaining all scores is not practical. To overcome this, instead of exhaustive search, we use POMDP-based DRL which observes a small number of scores (within optimized τ_{train}), and predicts the unobserved ones. In this way, the set of scores for all RF beams is obtained.

Table I shows the mapping of POMDP parameters to the parameters in (15). For fast convergence, both overestimation and underestimation of the value function should be avoided. We define a twin delayed deep deterministic policy gradient (TD3) agent $Q_{\pi_\phi}(s,a;\boldsymbol{\theta}_t^{\text{sel}},\boldsymbol{\theta}_t^{\text{eval}})$ as the function approximator for the optimal action $V_\pi^*(s,a)$, where $\boldsymbol{\theta}_t^{\text{sel}}$ and $\boldsymbol{\theta}_t^{\text{eval}}$ are TD3 weights at t used by the critic in DRL to select and evaluate a policy, respectively, and ϕ is the policy adopted by the actor in DRL. Two clipped Q-functions $Q_{\text{sel}}(s,a;\boldsymbol{\theta}_t^{\text{sel}})$ and $Q_{\text{eval}}(s,a;\boldsymbol{\theta}_t^{\text{eval}})$ are concurrently learned by minimizing the loss function (mean squared Bellman error (MSBE)).

Fig. 3 shows our POMDP-based DRL scheme for beamforming. The learning agent feeds its model with the observed $\mathbf{y}_{l,q,k,p}^{\text{Pilot}}[f]$ and the (incomplete) score table, as inputs and desired outputs, respectively. The aim is to learn the hidden

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

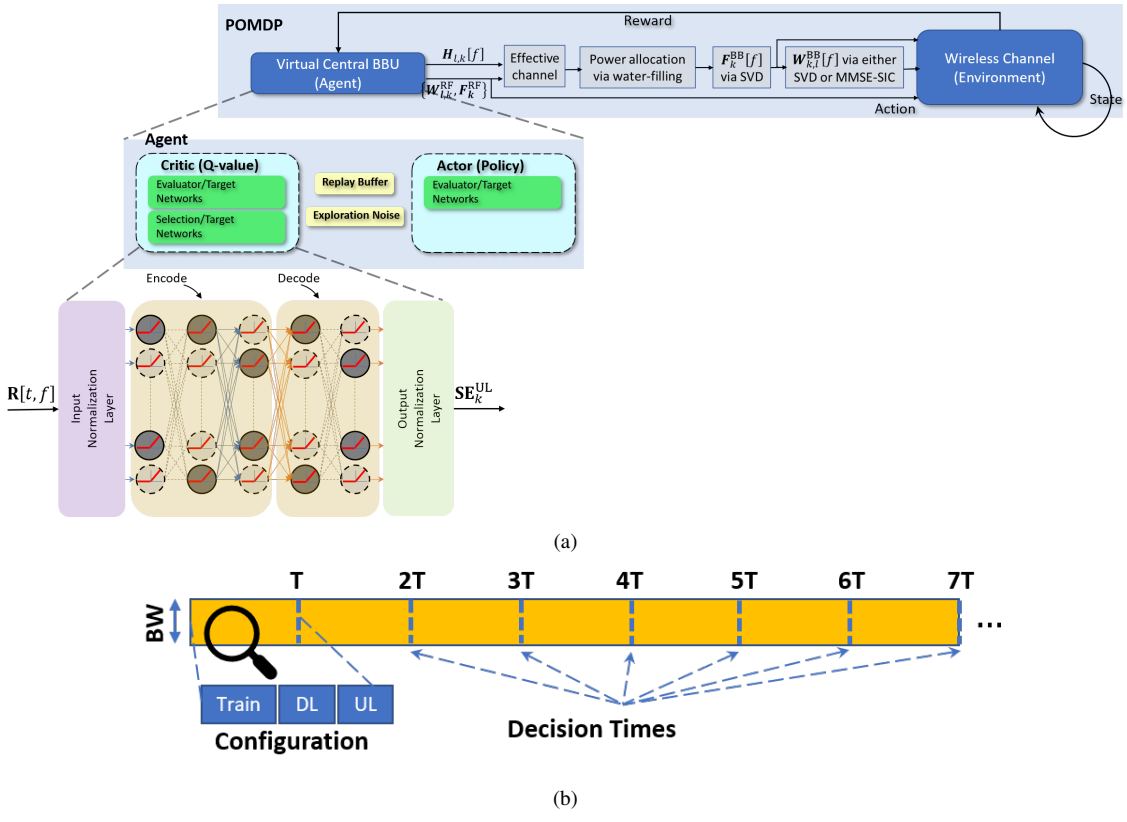


Figure 3. (a) A modular view of our DRL-based hybrid beamforming. (b) Timing: agent solves instances of (15) by adjusting the frame configuration.

relation between the jointly received signals by all APs and the rates of different sets of RF beams. Once trained, the agent can predict the best set of RF beams during the operation phase. The objective is to train the DRL's neural network, i.e., obtain the trained matrix of weight values that when multiplied by the channel vector for each instance and location, the beam is set to the desired direction. Beam steering is sequentially setting the beam's direction, depending on variations in channel vectors. The channel vector to each AP is impacted by fast mobility (i.e., varying from LoS to NLoS and blockage), which may be correlated to other APs' channels. The matrix of weight values can be considered as the concatenation of different vectors, where each vector corresponds to a different channel vector. We also present a low complexity training algorithm for our scheme, in which we directly map the received pilot signals into hybrid beamforming vectors to steer beams without excessive computations/signaling to obtain CSI.

In DRL, convergence of TD3 is not guaranteed [42], but a near-optimal policy can be found even when an arbitrary off-policy algorithm (exploratory policy) is used by the agent to select actions. We choose the ϵ-greedy policy to select actions, which is known to have a linear regret (in time) as shown in Algorithm 1. In doing so, we rewrite (17) as

$$y_{t+1} = r_t + \gamma \max_{\hat{a}} \Big[\lambda \min_{\boldsymbol{\theta}_t \in \{\boldsymbol{\theta}_t^{\text{eval}}, \boldsymbol{\theta}_t^{\text{sel}}\}} Q(s_{t+1}, \hat{a}; \boldsymbol{\theta}_t) + (1-\lambda) \max_{\boldsymbol{\theta}_t \in \{\boldsymbol{\theta}_t^{\text{eval}}, \boldsymbol{\theta}_t^{\text{sel}}\}} Q(s_{t+1}, \hat{a}; \boldsymbol{\theta}_t) \Big], \tag{20}$$

where λ is the weight of the minimum learned value, and

$$\hat{a} = \pi_{\phi_{t+1}}(s_t) + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \sigma^2), -c, c), \tag{21}$$

in which ϵ is the exploration noise clipped by c. As can be seen in simulations, our approach is stable, avoids convergence to local minima, and also avoids overfitting. Setting hyperparameter values in machine learning algorithms to minimize the learning error is a demanding task, but in what follows, we show that the learning error in TD3 is bounded.

It is well known from learning theory that generalization error is upperbounded [43]. Hence, the learning MSE of our proposed scheme, denoted by ξ, is upper bounded, i.e., $\xi < \frac{1}{2M}\left(2^{I(T_\xi;X)} + \log(1/\delta)\right)$, where δ is the confidence level, M is the number of training examples, X is TD3 input, T_ξ is the ξ-partition of X, and $I(T_\xi; X)$ is the mutual information of T_ξ and X, which depends on the neural network model [43], and can be used to choose a suitable model.

An advantage of our off-policy DRL is that an *a priori* deterministic target policy operates while the training (behavior) policy explores all possible beamforming actions by utilizing its own dataset. This, however, may introduce exploration bias [44] when the operation and training datasets are uncorrelated.

Table I
MAPPING POMDP PARAMETERS TO PARAMETERS IN (15)

Symbol	DRL Description	Our Problem
-	Agent	Virtual central baseband unit
$\mathcal{S}$	System state	Instantaneous channel state information
$\mathcal{A}$	Action set	Precoding and beamforming matrices
P	Environment	Stochastic uplink wireless channel
R	Reward function	QoS-aware EE^{UL}
Ω	Limited observations	SE_k^{UL} and EE^{UL}
O	Observation function	SE_k^{UL} and EE^{UL} estimates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

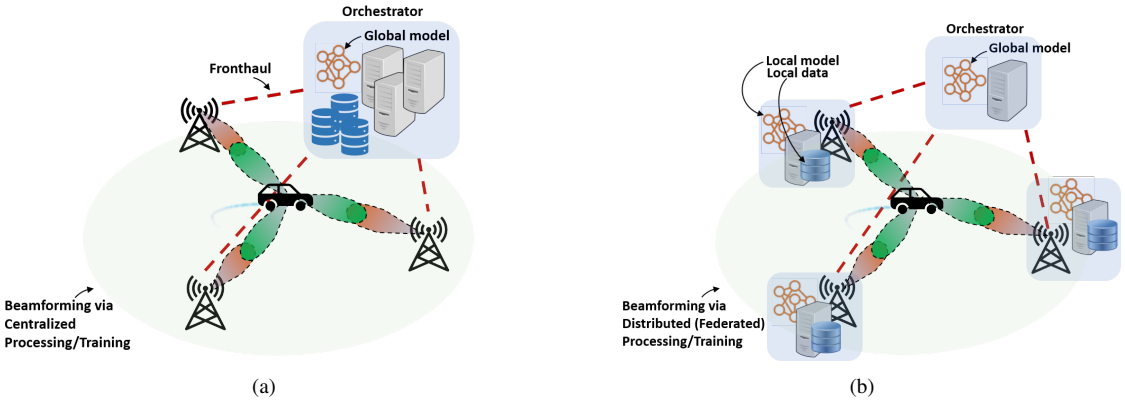


Figure 4. (a) Beamforming via centralized processing/training. (b) Beamforming via distributed (federated) processing/training.

To mitigate this bias, the operation and training datasets must be correlated but only to the extent that generalization is not impeded. In other words, they should be similarly distributed, called the coverage assumption in [38, Section 5]. In our scheme, we use ϵ-greedy policy, i.e., the operation and training datasets are correlated with a probability of $1-\epsilon$; and the training policy explores all actions with a probability ϵ.

Low-complexity training algorithms that are based on approximate message passing (AMP) are proposed in [45], [46] for single-layer networks, where damped AMP is used to improve stability at the cost of slow convergence. In Algorithm 2, we propose a low-complexity AMP-based training algorithm with unitary transformation for fast convergence. In case of large right-orthogonally invariant priors, Algorithm 2 is convergent and its convergence rate depends on the spectrum of the observation matrix. This can be easily shown by considering Algorithm 2 in terms of the factor graph, which is based on factorization of (16). Proof of convergence and its convergence rate follow directly from applying AMP to the factor graph (see Appendix A).

B. Baseband Beamforming

Once the RF beamforming matrices are obtained, i.e., when a connection is established and noise level is low, the baseband beamforming matrices $\mathbf{W}_{l,k}^{\mathrm{BB}}[f]$ and $\mathbf{F}_{k}^{\mathrm{BB}}[f]$ are constructed as the normalized left-singular vectors and right-singular vectors of their respective effective channel, i.e., $\mathbf{H}_{k,l}^{\text{effective}}[f] = \mathbf{W}_{l}^{\mathrm{RF}}\mathbf{H}_{k,l}[f]\mathbf{F}_{k}^{\mathrm{RF}}$, as widely reported in the literature. When noise level is high, MMSE-SIC receiver is used. Note that in baseband beamforming, both calculations and decisions are centralized, but in RF beamforming, only the decision is centralized. In certain cases, such as single-antenna single-user, distributed RF decisions are also optimal. The achievable rate is

$$\mathrm{SE}^{\mathrm{POMDP}} = \left(1 - \frac{|\mathcal{W}| \times |\mathcal{F}| \times T_p}{T}\right) \sum_{k=1}^{K} \mathrm{SE}_k^{\mathrm{UL}}. \quad (22)$$

Proposition 1. *Let SE^* denote the achievable rate when perfect CSI is available. Now, SE^{POMDP} almost surely converges to its upper bound SE^* as the number of antennas grows to infinity.*

Proof. The proof is straightforward and follows from maximizing the Rayleigh quotient in fully digital beamforming. □

Algorithm 1 QoS-Aware DRL-HBF

Require: $R_k \forall k, \mathcal{W}, \mathcal{F}$.

Ensure: Precoding and beamforming matrices are optimal.

Initialize: time slot $t=0$; critic networks $Q_{\theta_t^{\mathrm{sel}}}, Q_{\theta_t^{\mathrm{eval}}}$ and actor network π_{ϕ_t} with random numbers; target networks $\theta_{t+1}^{\mathrm{sel}} \leftarrow \theta_t^{\mathrm{sel}}$, $\theta_{t+1}^{\mathrm{eval}} \leftarrow \theta_t^{\mathrm{eval}}$, $\phi_{t+1} \leftarrow \phi_t$; replay buffer $\mathcal{B}$.
while TRUE **do**
  if MSBE$>\Delta$ for M steps & $t \bmod T_q \leq \tau_{\max}$ **then**

Training Phase: Agent learning

    Receive pilots to estimate $\hat{\mathbf{H}}_{l,k}$. ▷ Update state
    Restore ϵ to its initial value.
    Select action: $a_t \sim \pi_{\phi_t}(s_t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.
    Observe reward r_t and new state s_{t+1}.
    Store the experience (s_t, a_t, r_t, s_{t+1}) in $\mathcal{B}$.
    Sample mini-batch N transitions from $\mathcal{B}$.
    Smooth target policy $\hat{a}$ by (21).
    Update target network y_t by (20).
    Update critics by $\boldsymbol{\theta}_t \leftarrow \arg\min_{\boldsymbol{\theta}_t} \mathsf{MSE}(r_t - y_t)$
    if $t \bmod d$ **then**
      Update ϕ via deterministic policy gradient: $\nabla_\phi J_\phi = N^{-1} \sum \nabla_a Q_{\boldsymbol{\theta}_t}(s, a; \boldsymbol{\theta}_t)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
      Update target networks: $\boldsymbol{\theta}_{t+1}^{\mathrm{sel,eval}} \leftarrow \tau \boldsymbol{\theta}_t^{\mathrm{sel,eval}} + (1-\tau)\boldsymbol{\theta}_t^{\mathrm{sel,eval}}, \phi_{t+1} \leftarrow \tau\phi_t + (1-\tau)\phi_t$
    end if
  else

Operation Phase: Agent interacting with environment

    Sample $\zeta \sim \mathrm{Uniform}(0,1)$.
    if $\zeta \leq \epsilon$ **then** ▷ ϵ-greedy policy
      Select operation vector $\phi(t)$ at random.
    else
      Decrease $\epsilon : \epsilon \leftarrow \epsilon\tau$ where $\tau < 1$.
      Select $\phi(t) = \arg\max_a Q_{\pi_{\phi_t}}(s, a; \boldsymbol{\theta}_t^{\mathrm{sel}}, \boldsymbol{\theta}_t^{\mathrm{eval}})$.
    end if
    Beamform by (1) and (2). ▷ Carry out action
    Receive feedback r_t by (9). ▷ Observe reward
  end if
end while

Δ and M are subjective measures of convergence.
We use recursive least squares to implement the filters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

Algorithm 2 Multi-layer AMP-based learning algorithm

Require: Forward i.e., $\mathbf{G}_l^+(\mathbf{R}_{k,l-1}^+, \mathbf{R}_{k,l}^-, \mathbf{\Gamma}_{k,l-1}^+, \mathbf{\Gamma}_{k,l}^-)$ and reverse i.e., $\mathbf{G}_l^-(\mathbf{R}_{k,l-1}^-, \mathbf{R}_{k,l}^+, \mathbf{\Gamma}_{k,l-1}^-, \mathbf{\Gamma}_{k,l}^+)$ estimators

Set $\mathbf{R}_{0,l}^- = \mathbf{0}$ and $\mathbf{\Gamma}_{0,l}^- = \mathbf{0}, l = 1, \ldots, L-1$.
Initialize all layers weights with random numbers.
while MSBE $< \Delta$ **do**

Phase I: Inference

  while $\|\mathbf{R}_{k,l}^{\pm} - \mathbf{R}_{k-1,l}^{\pm}\| / \|\mathbf{R}_{k,l}^{\pm}\| \leq \epsilon_{\text{th}}$ **do**
    Apply known input, i.e., (4).
    for $l = 1, \ldots, L-1$ **do** ▷ Forward Pass
      $\hat{\mathbf{Z}}_{k,l}^+ = \rho \mathbf{G}_l^+(\cdot) + (1-\rho)\hat{\mathbf{Z}}_{k-1,l}^+$
      $\mathbf{\Lambda}_{k,l}^+ = [\partial \mathbf{G}_l^+(\cdot) / \partial \mathbf{R}_{k,l}^-]^{-1} \mathbf{\Gamma}_{k,l}^-$
      $\mathbf{\Gamma}_{k,l}^+ = \mathbf{\Lambda}_{k,l}^+ - \mathbf{\Gamma}_{k,l}^-$,
      $\mathbf{R}_{k,l}^+ = (\hat{\mathbf{Z}}_{k,l}^+ \mathbf{\Lambda}_{k,l}^+ - \mathbf{R}_{k,l}^- \mathbf{\Gamma}_{k,l}^-)(\mathbf{\Gamma}_{k,l}^+)^{-1}$
    end for
    Apply known output, i.e., pilot signals.
    for $l = L-1, \ldots, 1$ **do** ▷ Reverse Pass
      $\hat{\mathbf{Z}}_{k,l}^- = \rho \mathbf{G}_l^-(\cdot) + (1-\rho)\hat{\mathbf{Z}}_{k-1,l}^-$
      $\mathbf{\Lambda}_{k,l}^- = [\partial \mathbf{G}_l^-(\cdot) / \partial \mathbf{R}_{k,l}^+]^{-1} \mathbf{\Gamma}_{k,l}^+$
      $\mathbf{\Gamma}_{k,l}^- = \mathbf{\Lambda}_{k,l}^- - \mathbf{\Gamma}_{k,l}^+$,
      $\mathbf{R}_{k,l}^- = (\hat{\mathbf{Z}}_{k,l}^- \mathbf{\Lambda}_{k,l}^- - \mathbf{R}_{k,l}^+ \mathbf{\Gamma}_{k,l}^+)(\mathbf{\Gamma}_{k,l}^-)^{-1}$
    end for
  end while

Phase II: Tuning weights for each layer

  for all $\mathbf{z}_l, l = 1, \ldots L-1$ **do**
    Compute economy-sized SVD of $\mathbf{A}_l = \mathbf{U}\mathbf{S}\mathbf{V}^T$.
    Initialize $\mathbf{r}_1^0$ and γ_1^0.
    while $\|\mathbf{r}_1^t - \mathbf{r}_1^{t-1}\| / \|\mathbf{r}_1^t\| \leq \epsilon_{\text{th}}$ **do**
      $\hat{\mathbf{w}}_l^t = \rho \mathbf{g}_l(\mathbf{r}_r^t, \gamma_1^t) + (1-\rho)\hat{\mathbf{w}}_l^{t-1}$
      $\alpha_1^t = 1/N \sum_j \frac{\partial}{\partial r_j} \mathbf{g}_l(\mathbf{r}_r^t, \gamma_1^t)$
      $\mathbf{r}_2^t = \frac{1}{1-\alpha_1^t}(\hat{\mathbf{w}}_l^t - \alpha_1^t \mathbf{r}_1^t)$
      $\gamma_2^t = \gamma_1^t \frac{1-\alpha_1^t}{\alpha_1^t}$
      $\alpha_2^t = 1/N \sum_j \gamma_2^t / (s_j^2 / \hat{\tau}_w + \gamma_2^t)$
      $\mathbf{r}_1^{t+1} = \mathbf{r}_2^t + \frac{\mathbf{V}(\mathbf{S}^2 + \hat{\tau}_w \gamma_2^t \mathbf{I})^{-1} \mathbf{S}(\mathbf{U}^T \mathbf{z}_l - \mathbf{S}\mathbf{V}^T \mathbf{r}_2^t)}{1-\alpha_2^t}$
      $\gamma_1^{t+1} = \rho \gamma_2^t \frac{1-\alpha_2^t}{\alpha_2^t} + (1-\rho)\gamma_1^t$
    end while
  end for
end while

Latent variable of l-th layer is denoted by $\mathbf{Z}_l$. In training, $\mathbf{Z}_0$ and $\mathbf{Z}_L$ are known. $\mathbf{R}_{k,l}^+$ and $\mathbf{\Gamma}_{k,l}^+$: mean and precision (inverse variance) of the Gaussian messages in the forward direction; $\mathbf{R}_{k,l}^-$ and $\mathbf{\Gamma}_{k,l}^-$ represent the same quantities in the reverse direction.

C. Beamforming via Distributed Processing/Training

As shown in Fig. 4(a), in our centralized beamforming scheme, for each UE k, the APs in $\mathcal{M}_k$ use a virtual central processing unit to manage and learn beamforming with a view to connecting the UE to at least one AP. In centralized schemes, communications overhead as well as beam steering latency may be high. To reduce this overhead, we develop a scheme for beamforming via distributed processing/training shown in Fig. 4(b) by utilizing federated DRL in which APs in $\mathcal{M}_k$ participate in learning (training) and beamforming (operation) by way of distributed computing and centralized decision making. Specifically, each AP uses its locally processed data and collaborate with other APs in $\mathcal{M}_k$ to train a shared beamforming model orchestrated by the virtual central baseband unit.

In this scheme, each AP l in $\mathcal{M}_k$ uses its local data to iteratively train its neural network (which has the same architecture as that in the virtual central baseband unit) for steering beams. Local training is done via the scheme in Section IV-A. The weight values in each AP l (instead of local training data) are transmitted to the virtual central baseband unit for obtaining a set of shared weight values via aggregation. This significantly reduces the communications overhead. The virtual central baseband unit *aggregates* the received weight values into a shared set of weight values, which is transmitted back to all APs in $\mathcal{M}_k$. At convergence, the shared weight values will be the same as the local weight values.

We formulate the aggregation as a consensus optimization problem

$$\underset{\mathbf{w}_l : \mathbf{w}_l = \mathbf{z}}{\text{minimize}} \quad \sum_{l=1}^{L} f_l(\mathbf{w}_l), \tag{23}$$

where $f_l(\cdot)$ and $\mathbf{w}_l$ are the loss function and the vector of local model variables for AP l, respectively, and $\mathbf{z}$ is the vector of shared model variables. The constraints $\mathbf{w}_l = \mathbf{z}$ enforce consistency, or consensus. We solve (23) using Alternating Direction Method of Multipliers (ADMM). Each iteration of ADMM reduces to the following updates

$$\mathbf{w}_l^{t+1} = \underset{\mathbf{w}_l}{\text{argmin}} \left(f_l(\mathbf{w}_l) + (\rho/2) \left\| \mathbf{w}_l - \overline{\mathbf{w}}^t + \mathbf{u}_l^t \right\|_2^2 \right) \tag{24}$$

$$\mathbf{u}_l^{t+1} = \mathbf{u}_l^t + \mathbf{w}_l^{t+1} - \overline{\mathbf{w}}^{t+1} \tag{25}$$

where $\overline{\mathbf{w}}^t = \frac{1}{L} \sum_{l=1}^{L} \mathbf{w}_l^t$, ρ is the augmented Lagrangian parameter, and t is the iteration number.

V. Numerical Results and Discussion

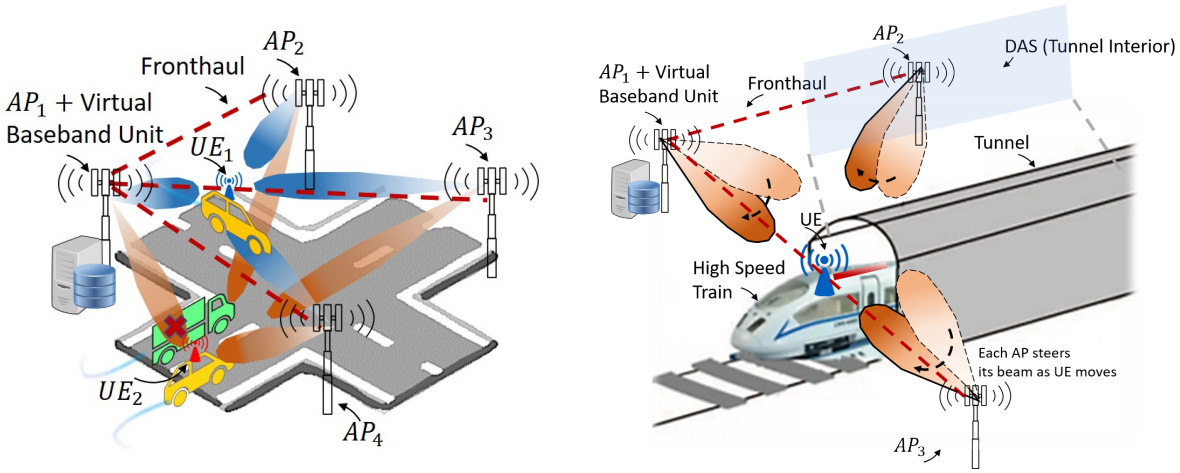
We numerically evaluate our schemes in two important high mobility use cases for urban and rural deployments, namely V2I and T2I. Our simulation method and rate evaluation are similar to [47]–[49]. Table II shows our simulation setup for each scenario, which corresponds to measurement-based channel models and specifications by 3rd generation partnership project (3GPP). Channel coefficients are generated as per the procedure in [33]. As shown in Fig. 5, for V2I, a passing truck blocks the LoS between the vehicle-under-test and the infrastructure; and for T2I, a HST with links to infrastructure exits a tunnel with semicircular cross-sections.

We model a two-part connection for end-users in each case. Part 1 involves AP, UE (not end-users), and outdoor channels, and Part 2 involves UE, end-users, and indoor channels. UE is mounted on the exterior (usually on top) of the vehicle/HST. Our focus is on Part 1, and we use Keras libraries with a TensorFlow backend for implementing DRL, and MATLAB® for baseband processing. A total of $K \times N_s$ RF chains are group-connected to the antenna via 10 fixed phase shifters [9]. Performance metrics are QoS-aware UL EE, convergence time, and communications overhead. We use time-domain uplink pilot signals (Zadoff-Chu sequences) for channel estimation, and benchmark our scheme against the conventional pilot-based scheme and an offline trained network where the optimal solution is obtained via exhaustive search. Figs. 6(a) to 6(h) show simulation results.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

Table II
SIMULATION SETUP

	Vehicular street-level application	**High-speed train**
System setup	$L = 4$ access points serving $K = 2$ UEs	$L = 3$ access points serving $K = 1$ UE
Access points	$N^{\text{AP-RF}} = 512, N^{\text{AP-BB}} = 1$	$N^{\text{AP-RF}} = 512, N^{\text{AP-BB}} = 1$
UEs	$P_{k,\max} = 30$ dBm, $R_k = 0, N_s = N^{\text{UE-BB}} = N^{\text{UE-RF}} = 4$, OFDM with 1024 sub-channels	$P_{k,\max} = 30$ dBm, $R_k = 0, N_s = N^{\text{UE-BB}} = N^{\text{UE-RF}} = 2$, OFDM with 1024 sub-channels
Channels	60 GHz band. For each channel realization, users are located in a 40m $\times$ 60m grid with 0.1 m resolution. System bandwidth is 1 GHz and noise figure is 5 dB.	28 GHz band. For each channel realization, user is located in a 400m $\times$ 600m grid with 1 m resolution. System bandwidth is 1 GHz and noise figure is 5 dB.
DRL model	Per-dataset input normalization and per-AP output normalization. Six fully connected layers, each with 512 nodes using ReLU activation units. Each layer feeds a drop-out regularization layer with 0.5% dropout rate. Training dataset has a maximum of 240,000 samples with a batch size of 100 samples. Also, $\lambda = 0.75$.	



(a) (b)

Figure 5. Simulation scenarios: (a) the outdoor urban environment in which a moving vehicle under test is communicating with ground infrastructure and its LoS is blocked by a passing truck, (b) the rural T2I scenario in a macro-and-relay layout exiting a tunnel with a semi-circle cross section equipped with distributed antenna system (DAS). For simplicity of the figure, UEs' beams are not shown.

As shown in Figs. 6(a) and 6(b), when a blockage occurs repeatedly at a given time and duration for a beam, our scheme learns the timing of such blockage and *proactively* anticipates it for that beam, but offline training scheme fails due to *bad* (non-representative) training data. When the state space is (very) large, there is little chance of *good* offline training irrespective of the computing power and/or storage space. Figs. 6(c) and 6(d) show the achievable data rate in our scheme and in the pilot-aided scheme in a LoS setting. In both schemes, the rate initially increases with the number of beams due to the beamforming gain. As the number of beams increases, the training overhead becomes dominant in the pilot-aided scheme, which causes significant rate drops at higher velocities. This is not the case in our scheme because the uplink training time is optimized in (15) for each interval T. Specifically, when the channel is affected due to a new situation (e.g., weather changes, new blockages, etc.), the learning agent spends more time on the exploration (i.e., $\tau_{\text{train}} \lessapprox T$). As the agent learns new situations, it generalizes its learning to cover all such cases and spends less time on the exploration, i.e., $\tau_{\text{train}} \gtrapprox 0$. Meanwhile, the agent adjusts τ_{train} to balance between system efficiency and training time via the term $(1 - \frac{\tau_{\text{train}}}{T})$ in (15b). Also, variations in the achievable rate due to changes in velocity are minor.

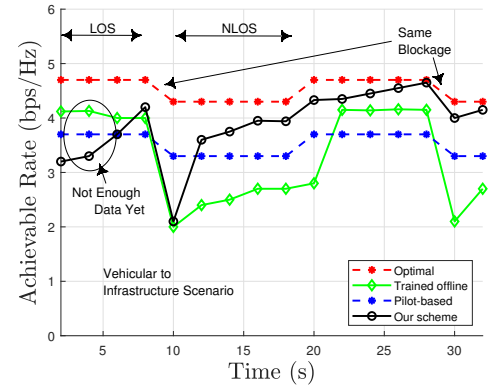
The 5th, 50th, and 95th percentile of cumulative distribution function (CDF) of SE for a given UE, which can be attributed to cell edge, median, and cell center UEs, respectively, are key performance indicators (KPIs) for service continuity, whose minimum required values are specified in IMT-2020 [55]. Fig. 6(e) shows that SE values for a fast-moving UE most likely varies from 1.5 bps/Hz (NLoS at cell-edge) to 4.2 bps/Hz (LoS at cell center). Also, SE values in Table III show that by using our hybrid beamformer, service remains available in spite of channel variations ranging from LoS to NLoS and blockage.

No scheme is *a priori* better than other schemes (including random beamforming) [56], but our scheme performs well by learning repeated irregularities in fast-moving UEs. As shown in Fig. 6(f), the *a posteriori* optimization method obtains the Pareto boundary in EE-SE plane that includes a critical operating point at which, energy efficiency peaks. Our scheme aims to reach this point in noise-limited paradigms via (15).

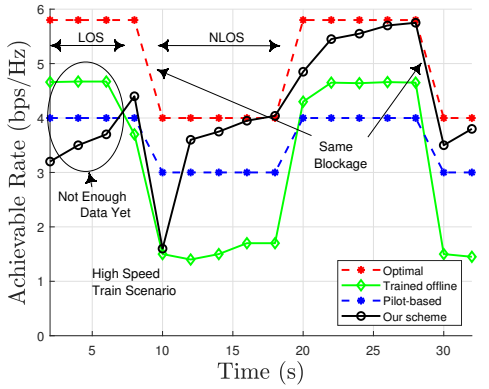
Table III
SE VALUES (IN BPS/HZ) IN UL FOR UE (R: IMT-2020, S: SIMULATION)

	V2I		T2I	
	R	S	R	S
5th %-tile of SE in UL for a UE	0.15	1.5	0.045	1.5
50th %-tile of SE in UL for a UE	5.4	6.2	1.6	3.1
95th %-tile of SE in UL for a UE	15	16.8	15	12.6

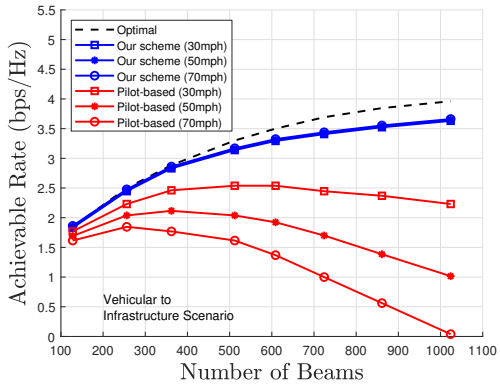
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications



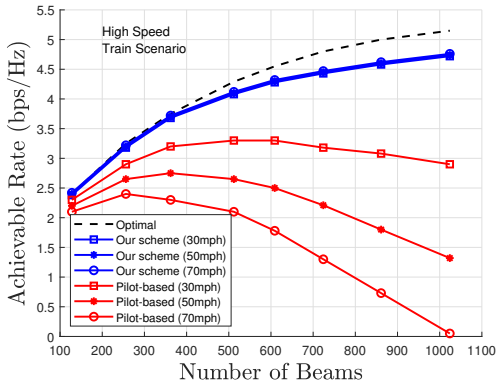
(a)



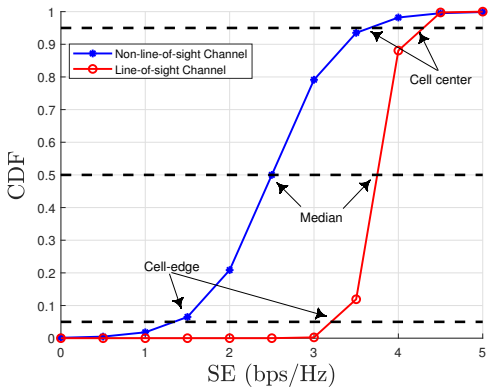
(b)



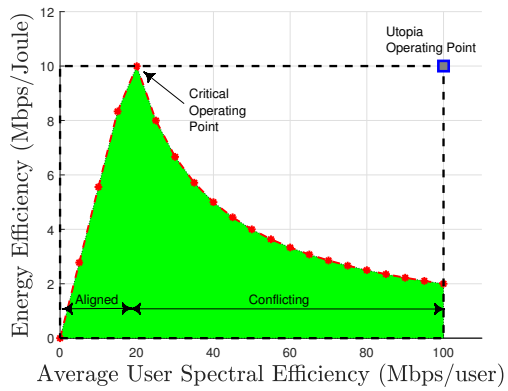
(c)



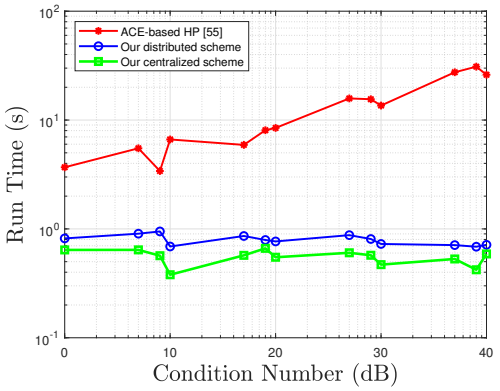
(d)



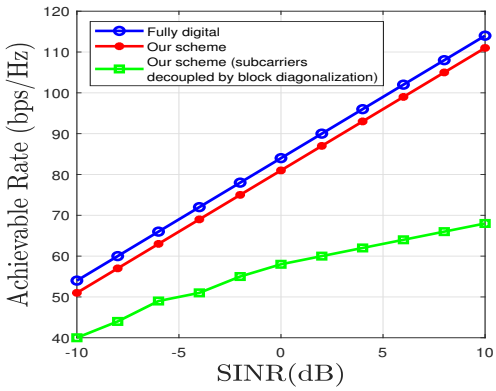
(e)



(f)



(g)



(h)

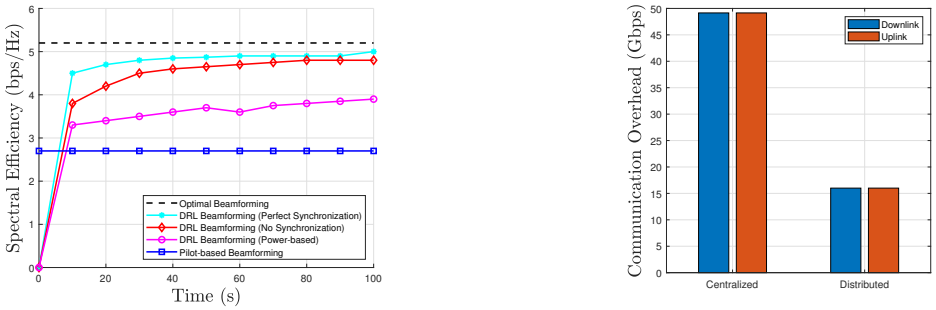
Figure 6. Simulation results: (a) and (b): Impact of blockage on achievable rates for V2I and T2I, respectively. (c) and (d): Impact of velocity on achievable rates for V2I and T2I, respectively. (e): Impact of blockage on CDF of a typical UE's achievable rates. (f): System Pareto boundary with critical operating point. (g): Typical run times of our schemes against other methods. (h): The effect of subcarrier decoupling on the achievable rate.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

Table IV
COMPUTATIONAL COMPLEXITY

No.	Method	Beamforming Time Complexity	Training Time Complexity	Run Time† (msec)
1	Exhaustive search	$\mathcal{O}(\lvert(\mathcal{U}\times\mathcal{F})^K\times\mathcal{W}^L\rvert)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	87540
2	OMP [10]	$\mathcal{O}(K^3F^3N_tN_{\text{RF}}^t(LN_s+{N_{\text{RF}}^t}^2+2N_{\text{RF}}^tN_s))$ ††	$\mathcal{O}(T_q N^{\text{AP-RF}})$	9839
3	PIS-MIB-SOMP [50]	$\mathcal{O}(K^3F^3N_tN_{\text{RF}}^t(LN_s+{N_{\text{RF}}^t}^2+2N_{\text{RF}}^tN_s))$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	6560
4	GS-HP [51]	$\mathcal{O}(K^3F^3N_tN_{\text{RF}}^t(LN_s+N_{\text{RF}}^t+2N_s))$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	1985
5	SIC-based HP [4]	$\mathcal{O}(K^3F^3((N_t/N_{\text{RF}}^t)^2(N_{\text{itr}}N_{\text{RF}}^t+N_r)+2N_{\text{RF}}^tN_{\text{itr}}))$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	1040
6	Element-wise [8]	$\mathcal{O}(K^3F^3N_{\text{itr}}N_t^4N_{\text{RF}}^t)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	1985
7	MO-AltMin [7]	Extremely high	$\mathcal{O}(T_q N^{\text{AP-RF}})$	43000
8	PE-AltMin [7]	$\mathcal{O}(N_{\text{itr}}{N_{\text{RF}}^t}^2N_tF^3K^3)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	850
9	SDR-AltMin [7]	$\mathcal{O}(N_{\text{itr}}{N_{\text{RF}}^t}^3N_s^3F^3K^3)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	1615
10	CR-MF [9]	$\mathcal{O}(N_{\text{RF}}^tN_sN_tKF)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	840
11	CR-MKM [9]	$\mathcal{O}(N_{\text{itr}}N_{\text{RF}}^tN_s^3F^3K^3)$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	240
12	FPS [52]	$\mathcal{O}(N_{\text{itr}}(K^2N_s^2N_{\text{RF}}^t+N_cN_{\text{RF}}^tN_t\log(N_cN_{\text{RF}}^tN_t)))$	$\mathcal{O}(T_q N^{\text{AP-RF}})$	125
13	DLHB [53]	$\mathcal{O}(\sum_{l=1}^{L_C}D_{x,l}D_{y,l}b_{x,l}b_{y,l}c_{C,l-1}c_{C,l})+\mathcal{O}(\sum_{l=1}^{L_F}b_{x,l}b_{y,l}c_{F,l})$††	-	1670
14	ACE-based HP [54]	$\mathcal{O}(N_{\text{itr}}\lvert\mathcal{W}\rvert N_tK^2)$	-	2430
15	Our scheme	$\mathcal{O}(\sum_{l=1}^{L_F}b_{x,l}b_{y,l}c_{F,l})$	-	120

† Numerical values of run time (also called wall time) are obtained via a server with 64 cores, where each core is an Intel(R) Core(TM) i5-7200U CPU @2.5GHz.
†† L, K, F and N_{itr} are the total number of paths in channels, the number of UEs, the number of subcarriers, and the number of iterations, respectively. Notations: L_C,L_F: number of convolutional and fully-connected layers; $D_{x,l}$,$D_{y,l}$: kernel dimensions; $b_{x,l}$,$b_{y,l}$: dimensions of the l-th convolutional layer output; $c_{C,l}$,$c_{F,l}$: number of filters in the l-th layer and units in the l-th fully-connected layer.



(a) (b)

Figure 7. (a) Spectral efficiency in our centralized processing/training scheme with different values of synchronization mismatch among APs in $\mathcal{M}_k$. (b) Comparison of the communication overhead in our centralized and distributed processing/training schemes in 5G NR.

We call the length of time for steering a beam during each channel coherence time T as the run time, which is the sum of operating and training times. During each T, training is iterative. Each iteration involves certain calculations (computational complexity (CC)) that depend on the implementation of backpropagation algorithm in the neural network, but operation is not iterative and its CC is straightforward. The number of iterations until convergence depends on the quality of MIMO channels. Note that (20) avoids over/underfitting the model, hence fewer iterations. MIMO channel quality is inversely related to channel variations and co-channel interference, and can be expressed by the condition number of channel matrix, defined as the ratio of the largest to the smallest singular values in singular value decomposition of that matrix. Offline training fails to steer beams in a timely manner due to significant channel variations for fast-moving UEs. Fig. 6(g) shows that run time in our schemes is stable in spite of channel variations and co-channel interference, and is significantly less than those of other online training schemes.

In hybrid transceivers, beamforming involves both analog and digital parts. The size of matrices in analog beamforming is much larger than that in digital beamforming, i.e., CC of hybrid beamformers is dominated by CC of the analog part. Existing methods typically trade off CC with hardware complexity and/or with spectral efficiency [57]. Note that CC of Algorithm 2 is dominated by singular value decomposition (SVD) of $\mathbf{A}_l$ in Step 21. The SVD of $\mathbf{A}\in\mathbb{R}^{L\times n}$ of rank r can be obtained with $\mathcal{O}(Lnr)$ floating-point operations (flops). Hence, CC of Algorithm 2 is similar to that of AMP algorithm.

Table IV compares CC of different schemes. In exhaustive search, all possible codebooks (the Cartesian product of transmit and receive analog beamforming matrices) are searched. For CC of methods 2-14 in Table IV, the interested reader is referred to the respective references. Our centralized and distributed schemes have the same order of polynomial CC, where in the latter, we move computations from the cloud into the edge to reduce communications overhead and steering latency.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3126056, IEEE Journal on Selected Areas in Communications

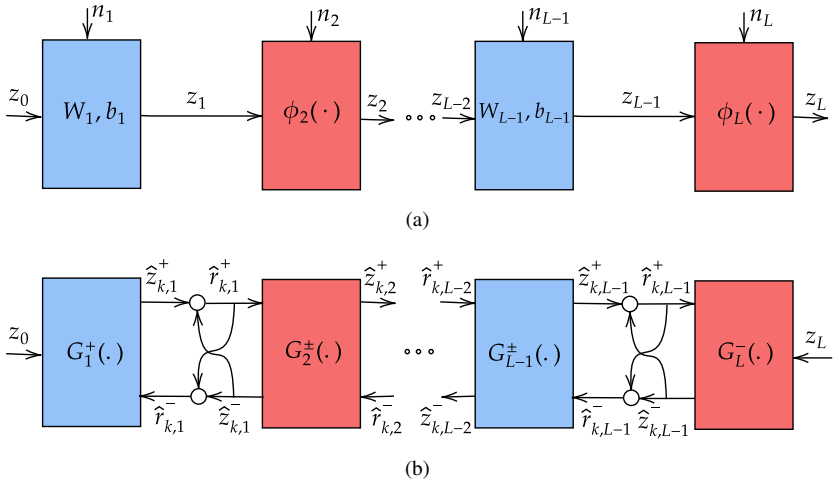


Figure 8. (a) Signal flow graph in POMDP-based DRL in Section IV-A. (b) Signal flow graph in Algorithm 2. Variables with superscript $+(-)$ are updated in the forward (backward) pass. Algorithm 2 solves (27) by sequentially solving a set of simpler estimation problems in consecutive pairs $(\mathbf{z}_l, \mathbf{z}_{l-1})$.

As stated in Section III-A, in wideband communications, one may utilize block diagonalization precoders to decouple subcarriers and reduce calculations. However, since RF beamformers are shared by many subcarriers and cause inter-user and inter-carrier interference, block diagonalization is ineffective and significantly reduces the achievable rate as shown in Fig. 6(h). Alternatively, filter bank multicarrier (FBMC) modulation that causes less out-of-band interference may be a better choice for fast-moving UEs despite its complexity.

Fig. 7(a) shows that our centralized scheme is not sensitive to phase synchronization among APs in $\mathcal{M}_k$. Fig. 7(b) shows that communication overhead in our distributed scheme (federated DRL) is significantly less than that of the centralized scheme, as the former only transmits the weight values of the neural network. Moreover, as stated earlier, our distributed scheme has the additional advantage of not requiring phase synchronization among APs in $\mathcal{M}_k$.

VI. Conclusion

In this paper, we utilized deep reinforcement learning for steering beams in a timely and efficient manner for high mobility communications in future networks. We also proposed a low-complexity federated DRL-based beamforming to significantly reduce communications overhead by utilizing edge computing instead of cloud computing. This significant reduction in the communications overhead has many practical advantages for fast and cost-effective deployment of 5G/6G networks in high mobility use cases. Simulation results demonstrate that our centralized and distributed processing/training schemes proactively learn to steer beams with superior performance as compared to other existing schemes.

Appendix A
On Convergence Analysis of Algorithm (2)

To analyze the convergence of Algorithm 2, consider an L-layer stochastic neural network, as in Fig. 3, given by

$$\mathbf{z}_l = \mathbf{W}_l \mathbf{z}_l + \mathbf{b}_l + \mathbf{n}_l, \qquad l = 1, 3, \ldots, L-1, \tag{26a}$$

$$\mathbf{z}_l = \phi_l(\mathbf{z}_{l-1}, \mathbf{n}_l), \qquad l = 2, 4, \ldots, L. \tag{26b}$$

where $\mathbf{z}_l \in \mathbb{R}^{d_l}$. The activation functions $\phi_l(\cdot)$ are nonlinear functions acting component-wise on their inputs. Also, $\mathbf{x}_{\text{train}} = \mathbf{z}_0$ and $\mathbf{y}_{\text{train}} = \mathbf{z}_L$ are the network's input and output, respectively. We are interested in the joint learning and inference problem under prior $\mathbb{P}(\mathbf{x}_{\text{train}}; \theta)$ and likelihood $\mathbb{L}(\mathbf{x}_{\text{train}}; \mathbf{y}_{\text{train}}, \theta)$, i.e.,

$$\text{Estimate} \qquad \{\mathbf{z}_l, \mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L-1} \tag{27a}$$

$$\text{Given} \qquad \{\mathbf{z}_0, \mathbf{z}_L\}. \tag{27b}$$

Consider the estimator in Fig. 8, based on insights from adaptive VAMP and multilayer VAMP algorithms. In what follows, we show that it is an asymptotically consistent estimator for (27), and analyze its convergence rate. Note that (26b) represents a Markov chain, hence the posterior $p(\mathbf{z}|\mathbf{z}_L)$ factorizes as $\mathbb{P}(\mathbf{y}, \mathbf{x}) = \mathbb{P}(\mathbf{x}) \mathcal{N}_{\mathbf{y}|\mathbf{A}_k\mathbf{x}}(\mathbf{0}, \gamma_w^{-1}\mathbf{I})$. (28)

Splitting $\mathbf{x}$ into two identical variables $\mathbf{x}_1 = \mathbf{x}_2$, gives an equivalent factorization

$$\mathbb{P}(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) = \mathbb{P}(\mathbf{x}_1)\delta(\mathbf{x}_1 - \mathbf{x}_2)\mathcal{N}_{\mathbf{y}|\mathbf{A}_k\mathbf{x}_2}(\mathbf{0}, \gamma_w^{-1}\mathbf{I}), \tag{29}$$

where $\delta(\cdot)$ is the Dirac delta distribution. This density can be represented as a linear factor graph with $L+1$ factors corresponding to $\mathbb{P}(z_0)$ and $\mathbb{P}(z_{l+1}|z_l)$, $l = 0, \ldots, L-1$. We consider both maximum a posteriori (MAP) and minimum mean squared error (MMSE) estimation for this posterior, i.e.,

$$\mathbf{z}_{\text{MAP}} = \arg\max_{\mathbf{z}} \mathbb{P}(\mathbf{z}|\mathbf{z}_0, \mathbf{z}_L) \tag{30a}$$

$$\mathbf{z}_{\text{MMSE}} = \mathbb{E}\{\mathbf{z}|\mathbf{z}_0, \mathbf{z}_L\} = \int \mathbf{z}\mathbb{P}(\mathbf{z}|\mathbf{z}_0, \mathbf{z}_L) d\mathbf{z}. \tag{30b}$$

Algorithm 2 produces estimates by a sequence of forward and backward pass updates for computing MAP and MMSE. We then pass messages on the corresponding factor graph according to the following rules:

1) Approximate beliefs: The approximate belief $b_{\text{app}}(\mathbf{x})$ on variable node $\mathbf{x}$ is $\mathcal{N}(\mathbf{x}|\hat{\mathbf{x}}, \eta^{-1}\mathbf{I})$, where $\hat{\mathbf{x}} = \mathbb{E}(\mathbf{x}|b_{\text{sp}})$ and $\eta^{-1} = \text{diag}(\text{Cov}(\mathbf{x}|b_{\text{sp}}))$ are the mean and average variance of the corresponding belief $b_{\text{sp}}(\mathbf{x}) = \prod_i \mu_{f_i \to \mathbf{x}}(\mathbf{x})$, i.e., is the normalized product of all messages impinging on the node.

2) Variable-to-factor messages: The message from a variable node $\mathbf{x}$ to a connected factor node f_i is $\mu_{\mathbf{x}\to f_i}(\mathbf{x}) = b_{\text{app}}(\mathbf{x})/\mu_{f_i\to\mathbf{x}}(\mathbf{x})$, i.e., is the ratio of the most recent approximate belief $b_{\text{app}}(\mathbf{x})$ to the most recent message from f_i to $\mathbf{x}$.
3) Factor-to-variable messages: The message from a factor node f to a connected variable node $\mathbf{x}_i$ is $\mu_{\mathbf{x}\to f_i}(\mathbf{x}) = \int f(\mathbf{x}_i, \{\mathbf{x}_j\}_{j\neq i}) \prod_{j\neq i} \mu_{\mathbf{x}_j\to f_i}(\mathbf{x}_j) d\mathbf{x}_j$.

By applying the above message-passing rules to the factor graph, the convergence of Algorithm 2 can be analyzed. The corresponding estimation functions are estimates of these belief densities.

Similar to [58], we analyze Algorithm 2 in a large system, where $d_0 \to \infty$ and $d_l, l \neq 0$ are fixed under rotatioanlly invariant random weight matrices, and obtain the corresponding *state evolution*, typically used to describe the mean-squared error of the estimates and test error. By concatenating the sequential estimates, we can think of Algorithm 2 as an equivalent adaptive VAMP algorithm for which rigorous convergence analysis exist, including state evolution equations, consistency of its fixed-point with that of the Bayes optimal estimator, and stability/sensitivity analysis [59]. Thus, the AMP-based Algorithm 2 provides a computationally tractable estimate with performance guarantees and testable conditions for optimality in certain high-dimensional random settings.

REFERENCES

[1] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.

[2] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, 2015.

[3] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, 2016.

[4] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, 2016.

[5] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—what is next? Five promising research directions for antenna arrays," *Digital Sig. Proc.*, 2019.

[6] X. Zhang, A. F. Molisch, and S.-Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE J. Sel. Topics Signal Process.*, vol. 53, no. 11, pp. 4091–4103, 2005.

[7] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, 2016.

[8] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, 2017.

[9] X. Yu, J. Zhang, and K. B. Letaief, "A hardware-efficient analog network structure for hybrid precoding in millimeter wave systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 282–297, 2018.

[10] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, 2014.

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

[12] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, 2016.

[13] P. Sudarshan, N. B. Mehta, A. F. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, 2006.

[14] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—the large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[15] 3GPP, "5G; Study on Scenarios and Requirements for Next Generation Access Technologies," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, 05 2017, version 14.2.0.

[16] M. Zhang, M. Polese, M. Mezzavilla, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "Will TCP work in mmWave 5G cellular networks?" *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 65–71, 2019.

[17] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018.

[18] A. Alkhateeb, I. Beltagy, and S. Alex, "Machine learning for reliable mmWave systems: Blockage prediction and proactive handoff," in *IEEE Global Conf. Signal and Info. Processing (GlobalSIP)*, 2018, pp. 1055–1059.

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

[20] X. Li, A. Alkhateeb, and C. Tepedelenlioğlu, "Generative adversarial estimation of channel covariance in vehicular millimeter wave systems," in *IEEE 52nd Asilomar Conf. Signals, Systems, and Computers*, 2018, pp. 1572–1576.

[21] Y. Guo, Z. Wang, M. Li, and Q. Liu, "Machine learning based mmWave channel tracking in vehicular scenario," in *IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2019, pp. 1–6.

[22] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, 2020.

[23] Y. Wang, M. Narasimha, and R. W. Heath, "MmWave beam prediction with situational awareness: A machine learning approach," in *IEEE 19th Inter. Workshop on Signal Proc. Advances in Wireless Commun. (SPAWC)*, 2018, pp. 1–5.

[24] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137 184–137 206, 2019.

[25] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, M. Morikura, and K. Nakashima, "Communication-efficient multimodal split learning for mmWave received power prediction," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1284–1288, 2020.

[26] V. Raghavan, L. Akhoondzadeh-Asl, V. Podshivalov, J. Hulten, M. A. Tassoudji, O. H. Koymen, A. Sampath, and J. Li, "Statistical blockage modeling and robustness of beamforming in millimeter-wave systems," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 3010–3024, 2019.

[27] Q. Wang, K. Feng, X. Li, and S. Jin, "PrecoderNet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Commun. Letters*, vol. 9, no. 10, pp. 1677–1681, 2020.

[28] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, "Performance optimization in mobile-edge computing via deep reinforcement learning," in *IEEE 88th Vehicular Technol. Conf. (VTC-Fall)*. IEEE, 2018, pp. 1–6.

[29] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[30] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[31] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 18–28, 2004.

[32] C.-X. Wang, J. Bian, J. Sun, W. Zhang, and M. Zhang, "A survey of 5G channel measurements and models," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3142–3168, 2018.

[33] Y. Liu, C.-X. Wang, J. Huang, J. Sun, and W. Zhang, "Novel 3-D nonstationary mmWave massive MIMO channel models for 5G high-speed train wireless communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2077–2086, 2018.

[34] 3GPP, "5G; Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, 05 2017, version 14.0.0.

[35] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4042–4058, 2017.

[36] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, 2016.

[37] L. Liu, C. Tao, J. Qiu, H. Chen, L. Yu, W. Dong, and Y. Yuan, "Position-based modeling for wireless channel on high-speed railway under a viaduct at 2.35 GHz," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 834–845, 2012.

[38] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*. MIT press, 2018.

[39] N. David, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple-access channels." *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1859–1874, 2004.

[40] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[41] K. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured Grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, 2020.

[42] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *35th International Conf. Machine Learning*, vol. 80. PMLR, 10–15 Jul 2018, pp. 1587–1596.

[43] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint: 1703.00810*, 2017.

[44] M. V. Pogančić. (2019) The false promise of off-policy reinforcement learning algorithms. [Online]. Available: https://towardsdatascience.com/the-false-promise-of-off-policy-reinforcement-learning-algorithms-c56db1b4c79a

[45] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[46] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[47] W. Zheng, A. Ali, N. González-Prelcic, R. Heath, A. Klautau, and E. M. Pari, "5G V2X communication at millimeter wave: rate maps and use cases," in *IEEE 91st Veh. Technol. Conf. (VTC2020-Spring)*, 2020, pp. 1–5.

[48] K. Guan, B. Ai, B. Peng, D. He, G. Li, J. Yang, Z. Zhong, and T. Kürner, "Towards realistic high-speed train channels at 5G millimeter-wave band—part I: paradigm, significance analysis, and scenario reconstruction," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9112–9128, 2018.

[49] ——, "Towards realistic high-speed train channels at 5G millimeter-wave band—part II: Case study for paradigm implementation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9129–9144, 2018.

[50] Y.-Y. Lee, C.-H. Wang, and Y.-H. Huang, "A hybrid RF/baseband precoding processor based on parallel-index-selection matrix-inversion-bypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 305–317, 2014.

[51] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, 2016.

[52] X. Yu, J. Zhang, and K. B. Letaief, "Hybrid precoding in millimeter wave systems: How many phase shifters are needed?" in *IEEE Global Commun. Conf. (GLOBECOM)*, 2017, pp. 1–6.

[53] A. M. Elbir, K. V. Mishra, M. Shankar, and B. Ottersten, "Online and offline deep learning strategies for channel estimation and hybrid beamforming in multi-carrier mm-wave massive MIMO systems," *arXiv preprint arXiv:1912.10036*, 2019.

[54] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–6.

[55] M. Series, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," *Recommendation ITU*, vol. 2083, 2017.

[56] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[57] J. Zhang, X. Yu, and K. B. Letaief, "Hybrid beamforming for 5G and beyond millimeter-wave systems: A holistic view," *IEEE Open Journal of the Commun. Society*, vol. 1, pp. 77–91, 2019.

[58] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher, "Inference in multi-layer networks with matrix-valued unknowns," *arXiv preprint arXiv:2001.09396*, 2020.

[59] A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems," *Advances in Neural Information Processing Systems*, vol. 30, pp. 2452–2551, 2017.

Mahdi Fozi is a Ph.D. candidate in the Faculty of Electrical and Computer Engineering at Tarbiat Modares University, Tehran, Iran. He received his B.Sc. degree from University of Tabriz in Tabriz, Iran and his M.Sc. degree from Sharif University of Technology in Tehran, Iran, both in Electrical Engineering in 2012 and 2014, respectively. His research interests are advanced signal processing techniques, communications systems and networks, and artificial intelligence in future networks.

Ahmad R. Sharafat is a Professor of Electrical and Computer Engineering at Tarbiat Modares University in Tehran, Iran; Member of Iranian Academy of Sciences; Chairman of ITU-D Study Group 2 in the International Telecommunication Union (ITU) in Geneva, Switzerland; Past Chairman of IEEE Iran Section; Editor of the International Journal of Wireless Information Networks; and Editor of Scientia Iranica. He has 12 patents, co-authored 4 books and more than 150 papers in refereed scholarly journals and professional conferences. His research interests are advanced signal processing techniques, and communications systems and networks. He received his B.Sc. degree from Sharif University of Technology, Tehran, Iran, and his M.Sc. and his Ph.D. degrees both from Stanford University, Stanford, California, all in Electrical Engineering in 1975, 1976, and 1981, respectively. He is a Life Senior Member of IEEE and Sigma Xi.

Mehdi Bennis is a tenured full Professor at the Centre for Wireless Communications, University of Oulu, Finland, Academy of Finland Research Fellow and head of the intelligent connectivity and networks/systems group (ICON). His main research interests are in radio resource management, heterogeneous networks, game theory and distributed machine learning in 5G networks and beyond. He has published more than 200 research papers in international conferences, journals and book chapters. He has been the 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 of Wireless Communications and Networks, 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. Dr. Bennis is an editor of IEEE TCOM and Specialty Chief Editor for Data Science for Communications in the Frontiers in Communications and Networks journal. Dr. Bennis is an IEEE Fellow.