IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2020, accepted July 23, 2020, date of publication July 27, 2020, date of current version August 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012199

Energy-Efficient Access-Point Sleep-Mode Techniques for Cell-Free mmWave Massive MIMO Networks With Non-Uniform Spatial Traffic Density

JAN GARCÍA-MORALES, (Member, IEEE), GUILLEM FEMENIAS, (Senior Member, IEEE), AND FELIP RIERA-PALOU, (Senior Member, IEEE)

Mobile Communications Group, University of the Balearic Islands, 07122 Palma, Spain

Corresponding author: Guillem Femenias (guillem.femenias@uib.es)

This work was supported in part by the Agencia Estatal de Investigación and the Fondo Europeo de Desarrollo Regional (AEI/FEDER, UE), and in part by the Ministerio de Economía y Competitividad (MINECO), Spain, through the Project TERESA, under Grant TEC2017-90093-C3-3-R.

ABSTRACT Cell-free massive multiple-input multiple-output (MIMO) is a novel beyond 5G (B5G) and 6G paradigm that, through the use of a common central processing unit (CPU), coordinates a large number of distributed access points (APs) to coherently serve mobile stations (MSs) on the same time/frequency resource. By exploiting the characteristics of new less-congested millimeter wave (mmWave) frequency bands, these networks can improve the overall system spectral and energy efficiencies by using low-complexity hybrid precoders/decoders. For this purpose, the system must be correctly dimensioned to provide the required quality of service (QoS) to MSs under different traffic load conditions. However, only heavy traffic load conditions are usually taken into account when analysing these networks and, thus, many APs might be underutilized during low traffic load periods, leading to an inefficient use of resources and waste of energy. Aiming at the implementation of energy-efficient AP switch on/off strategies, several approaches have been proposed in the literature that only consider rather unrealistic uniform spatial traffic distribution in the whole coverage area. Unlike prior works, this paper proposes energy efficient AP sleep-mode techniques for cell-free mmWave massive MIMO networks that are able to capture the inhomogeneous nature of spatial traffic distribution in realistic wireless networks. The proposed framework considers, analyzes and compares different AP switch ON-OFF (ASO) strategies that, based on the use of goodness-of-fit (GoF) tests, are specifically designed to dynamically turn on/off APs to adapt to both the number and the statistical distribution of MSs in the network. Numerical results show that the use of properly designed GoF-based ASO strategies under a non-uniform spatial traffic distribution can serve to considerably improve the achievable energy efficiency.

INDEX TERMS Cell-free massive MIMO, energy efficiency, access-point switch on/off techniques, millimeter-wave communications, goodness-of-fit.

I. INTRODUCTION

A. MOTIVATION AND PREVIOUS WORK

Cell-free massive multiple-input multiple-output (MIMO), originally proposed by Ngo *et al.* in [1], is a novel wireless networking paradigm currently being investigated in the context of beyond 5G (B5G) and sixth generation (6G) mobile communications. One of the defining and foremost important features of cell-free massive MIMO is the replacement of the classical cell-based structure, pervasive in current wireless networks, by a large number of randomly deployed access points (APs) scattered throughout the coverage area, all connected to a central processing unit (CPU) that coordinates the communications [2]. Note that this architecture

The associate editor coordinating the review of this manuscript and approving it for publication was Yan Huo.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

eliminates the concept of cell altogether, hence also avoiding cell-edge performance issues. Importantly, the large number of APs, each potentially equipped with multiple antennas, makes cell-free massive MIMO a distributed form of the *classical* centralized massive MIMO, thus inheriting many of its attractive features such as the channel hardening and favourable propagation effects through the implementation of simple signal processing at both transmission ends. Recent research has shown that bringing the radio frequency (RF) front-end closer to the users while allowing certain operations to be conducted centrally (i.e., power and pilot allocation) significantly outperforms conventional architectures such as small cells or centralized MIMO strategies in providing a uniform quality of service (QoS) throughout the network [3], [4]. Moreover, the cell-free massive MIMO paradigm allows for a variety of trade-offs in terms of performance, complexity and fronthaul link requirements to be implemented depending on the capabilities of both the CPU and the APs and/or the capacities of the fronthaul links connecting the APs to the CPU. In particular, it was shown in [4], [5] that conducting the precoding operation at the CPU in a cooperative centralized manner while relying on instantaneous channel state information (CSI) leads to a considerable improvement over simpler AP-based non-cooperative beamforming strategies.

On another front, the rapid increase in mobile data demand over the last decade has virtually filled up most of the conventional mobile/wireless frequency bands (from 300 MHz to 6 GHz), leading to the so-called spectrum crunch and the need to explore alternative frequency regions to accommodate future standards/services [6]–[8]. The so-called millimeter wave (mmWave) band, located between 6 and 300 GHz, has emerged as the most viable candidate in the short term and in fact it is already playing a significant role in the roll-out of fifth generation (5G) networks [9], a trend likely to be fully developed in the context of B5G and 6G systems. Despite the large chunks of available spectrum at mmWave bands, this portion of spectrum presents important challenges from a communication point of view, most notably, a very severe propagation pathloss. Large antenna arrays with many radiating elements can be used to effectively implement mmWave massive MIMO schemes that, with appropriate beamforming, compensate for the orders-of-magnitude increase in free-space path-loss when compared to a sub-6 GHz environment (see [10], [11] and references therein). However, unlike sub-6 GHz communications, where processing is completely performed in the digital domain, at mmWave frequencies the energy consumption and hardware cost associated to the use of a large number of antenna elements precludes this possibility. Instead, use is typically made of hybrid digital-analog architectures whereby a large number of antennas is interfaced through an analog front-end (implemented using phase shifters) to a much smaller number of RF chains that take care of down-mixing and analog-to-digital conversion [12], [13]. The application of cell-free designs at mmWave was first considered in [14] (subsequently expanded in [15]) in the context of uncorrelated channels and mainly targeting power allocation strategies for energy-efficiency maximization. The effects of fronthaul limitations in a cell-free massive MIMO at mmWave frequencies under correlated fading was studied in [16] along with a proposal to conduct user selection in the likely event that the number of users in the system exceeds the number of RF chains at the APs.

The deployment of a large number of antennas, either in a centralized form as in massive MIMO or a distributed one as in cell-free massive MIMO, raises concerns regarding energy consumption both from the point of view of energy efficiency but also when considering the overall emissions produced by the mobile communication industry. In particular, the whole of information and communication technology (ICT) industry has been estimated to contribute up to about 23% of the global carbon footprint and about 51% of global energy electricity consumption by 2030 [17]. Addressing this issue, energy efficient wireless communication (also called *green communication*) has been an important research thread for well over a decade and it is envisaged to continue to do so for the foreseeable future [18]. Among the many *green* strategies that have been proposed, one that is specially effective in reducing the carbon footprint associated to cellular networks is the one based on the so-called *switch on/off algorithms* (see, for instance, [18]–[20] and references therein). Since most networks are designed and deployed to cope with fully loaded scenarios, a situation that most often is not sustained at all times, these techniques aim at dynamically turning on/off a fraction of the base stations (BSs) in response to variations in the user locations and traffic demands. Typically, these strategies are combined with specific forms of user association and cell zooming (i.e., cell breathing) see [21]–[24]) that boost their performance. Very recent research in [25] has shown that the application of *switch on/off* algorithms in the context of cell-free massive MIMO, whereby APs are dynamically (de)activated, has proved effective in optimizing the energy efficiency of the network. However, this work, which was framed in a sub-6 GHz context, relied on the assumption of having homogenously distributed users in the spatial domain, a condition seldom met in practice.

B. CONTRIBUTIONS

Motivated by the above works and open issues, our main goal in this paper is to propose a green networking solution encompassing the design, performance evaluation and comparison of energy-efficient access point switching (ASO) strategies for cell-free mmWave massive MIMO networks when considering non-uniform spatial traffic densities. In particular, the noteworthy contributions of our work can be summarized as follows:

- Tractable expressions are derived for the energy/spectral efficiencies, and the power consumption in cell-free mmWave massive MIMO networks for the particular case of zero-forcing (ZF) precoding/combining for both the downlink (DL) and uplink (UL). Remarkably, these mathematical expressions are able to account for the fact

that there is a limited number of active RF chains at each of the APs in the network.

- In contrast to previous works on switch on/off algorithms for cell-free massive MIMO networks (see, for instance, [25], [26]), this study contemplates a realistic model to describe a non-uniform distribution of mobile stations (MSs), thus capturing the heterogeneous nature of spatial traffic density in practical wireless networks [27].
- Aside from recasting existing adaptive ASO techniques to the scenario at hand, novel ASO strategies are implemented that rely on statistical goodness-of-fit (GoF) tests, and govern the process of (de)activating APs in such a way that the distribution of the resulting active AP matches the non-uniform spatial distribution of MSs. These GoF-based strategies are shown to be computationally simple and to only depend on information regarding the location of APs and the long-term spatial distribution of MSs. In comparison to previous techniques, our proposals attain a much better trade-off in terms of jointly assessing performance, complexity and implementability.
- Extensive simulation results show the energy efficiency benefits of the proposed practical ASO strategies under a comprehensive set of cell-free mmWave massive MIMO scenarios. In particular, the impact the number of MSs in the network and the RF infrastructure used at the APs have on the spectral/energy efficiency of the proposed network is evaluated under various non-uniform spatial distributions of MSs.

C. PAPER ORGANIZATION AND NOTATIONAL REMARKS

The proposed green cell-free mmWave massive MIMO network is introduced in Section II, where different subsections are dedicated to describe the spatial modeling of the distribution of MSs, the spatially correlated channel model at mmWave bands, the RF precoder/decoder design, the algorithm used to select the MSs to beamform from each active AP, the UL training phase and, finally, the DL and UL payload transmission phases. The different performance metrics used in this paper, including the spectral efficiency, the power consumption model and the energy efficiency, are thoroughly evaluated in Section III. The proposed ASO strategies, assuming scenarios with non-uniform spatial traffic distribution, are fully described in Section IV. Numerical results and discussions are provided in Section V and, finally, Section VI concludes the paper.

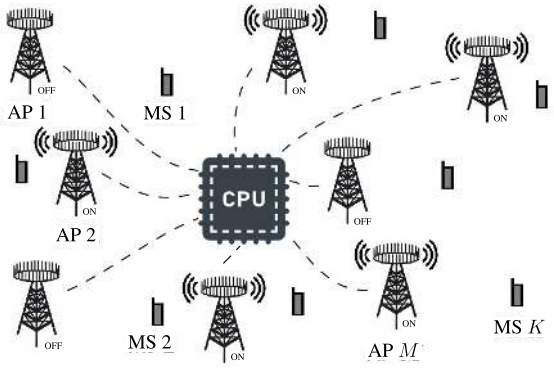
Notation: Vectors and matrices are denoted by lower- and upper-case boldface symbols, respectively. The q-dimensional identity matrix is represented by $\boldsymbol{I}_q$. The operator $\|\boldsymbol{x}\|$ represents the Euclidian norm of vector $\boldsymbol{x}$, whereas $\boldsymbol{X}^{-1}$, $\boldsymbol{X}^T$, $\boldsymbol{X}^*$ and $\boldsymbol{X}^H$ denote the inverse, transpose, conjugate and conjugate transpose (also known as Hermitian) of matrix $\boldsymbol{X}$, respectively. With a slight abuse of notation, the operator diag($\boldsymbol{x}$) is used to denote a diagonal matrix with the entries of vector $\boldsymbol{x}$ on its main diagonal, whereas diag($\boldsymbol{X}$) is used to denote a vector containing the elements of the main diagonal of matrix $\boldsymbol{X}$. The expectation operator is denoted by $\mathbb{E}\{\cdot\}$. Finally, $\mathcal{CN}(\boldsymbol{m}, \boldsymbol{R})$ denotes a complex Gaussian vector distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{R}$, $\mathcal{N}(0, \sigma^2)$ denotes a real valued zero-mean Gaussian random variable with standard deviation σ, and $\mathcal{U}[a, b]$ represents a random variable uniformly distributed in the range $[a, b]$. In order to ease the reading of this paper, Table 1 summarizes the definition of some of the most important parameters and variables used in the different subsections (see also Table 2 summarizing the default simulation parameters used in the numerical results section).

TABLE 1. Summary of main parameters and variables.

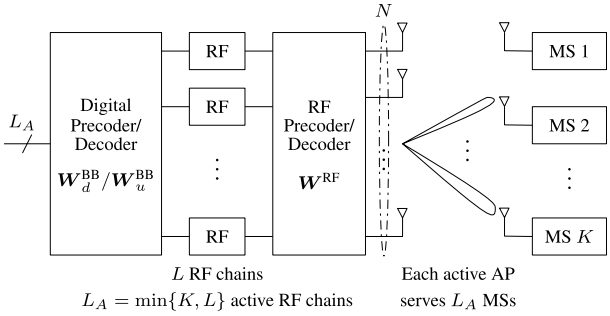
Parameter	Description
M	Number of APs
M_A	Number of active APs
N	Number of antennas at each AP
L	Number of available RF chains at each AP
L	Number of aactive RF chains at each AP
K	Number of MSs
$\check{\boldsymbol{h}}_{mk}$	Channel between the kth MS and the mth AP
$\overline{\boldsymbol{h}}_{mk}$	Normalized LOS component of $\check{\boldsymbol{h}}_{mk}$
$\boldsymbol{h}_{mk}$	Normalized NLOS component of $\check{\boldsymbol{h}}_{mk}$
$\check{R}_{mk}$	Spatial covariance matrix of $\check{\boldsymbol{h}}_{mk}$
R_{mk}	Spatial covariance matrix of $\boldsymbol{h}_{mk}$
$\boldsymbol{W}_m^{\mathrm{RF}}$	Analog RF precoder/combiner stage at the mth AP
$\boldsymbol{W}_{d\,m}^{\mathrm{BB}}/\boldsymbol{W}_{u\,m}^{\mathrm{BB}}$	Digital precoder/combiner stage at the mth AP
$\boldsymbol{g}_{mk}$	Equivalent channel (including RF precoder/decoder) between the kth MS and the mth AP
$\hat{\boldsymbol{g}}_{mk}$	MMSE estimation of $\boldsymbol{g}_{mk}$
$\check{\boldsymbol{H}}_m$	MIMO channel between the mth AP and the K MSs
$\boldsymbol{G}_m$	Equivalent MIMO channel (including RF precoder/decoder) between the mth AP and the K MSs
$S_{ed}(\boldsymbol{\upsilon})/S_{eu}(\boldsymbol{\omega})$	DL/UL spectral efficiency as a function of the vector of DL/UL
$P_{Td}(\boldsymbol{\upsilon})/P_{Tu}(\boldsymbol{\omega})$	DL/UL power consumption as a function of the vector of DL/UL power control coefficients
$E_{ed}(\boldsymbol{\upsilon})/E_{eu}(\boldsymbol{\omega})$	DL/UL energy efficiency as a function of the vector of DL/UL power control coefficients
$E_e(\boldsymbol{\upsilon}, \boldsymbol{\omega})$	Weighted energy efficiency as a function of the vector of DL and UL power control coefficients

II. SYSTEM MODEL

As shown in Fig. 1, we consider a cell-free mmWave massive MIMO network where M randomly distributed APs, each equipped with an array of N antennas and connected to a CPU via an infinite-capacity error-free fronthaul link, can be activated to provide service to K single-antenna MSs. The process of activation/deactivation of APs is basically driven by the implementation of ASO strategies, with APs in active (ON) and sleep (OFF) modes being indexed by the sets $\mathcal{M}^A = \{m_1^A, \ldots, m_{M_A}^A\}$ and $\mathcal{M}^S = \{m_1^S, \ldots, m_{M_S}^S\}$, respectively, where $\mathcal{M}^A \cap \mathcal{M}^S = \emptyset$ and $\mathcal{M}^A \cup \mathcal{M}^S = \{1, \ldots, M\}$. Moreover, as the implementation of a dedicated RF chain to each antenna results in unaffordable energy consumption and



(a) Cell-free mmWave massive MIMO network where M_A active APs (out of M available APs) provide service to K MSs.



(b) Generic transmitter/receiver architecture: hybrid analog-digital precoder/decoder with L RF chains, and fully connected phase shifter network.

FIGURE 1. System model of a cell-free mmWave massive MIMO network using ASO strategies and hybrid analog-digital precoding/decoding schemes.

hardware cost in mmWave-based massive MIMO systems, the number of required RF chains is reduced by relying on the use of hybrid analog-digital precoding schemes. In particular, as shown in Fig. 1b, it is assumed in this paper that each AP is equipped with $L \leq N$ RF chains and that a fully-connected architecture is arranged where each RF chain is connected to all the antenna elements using N analog phase shifters.

The communication between the active APs and the MSs is coordinated by the CPU through the use of a half-duplex time division duplexing (TDD) algorithm in which each frame is divided into three phases, namely, the UL training phase, the UL payload data transmission phase and the DL payload data transmission phase. During the UL training phase, all MSs transmit training pilots allowing the active APs to estimate the propagation channels to every MS in the network. Channel estimates are then used to detect the signals transmitted from the MSs in the UL payload data transmission phase and to compute the precoding filters governing the DL payload data transmission. The combined duration/bandwidth of the training, UL and DL phases, denoted as τ_p, τ_u and τ_d, respectively, should not exceed the coherence time/bandwidth of the channel, denoted as τ_c, that is, $\tau_p + \tau_d + \tau_u \leq \tau_c$, with all these intervals specified in samples (or channel uses) on a time-frequency grid. It is worth pointing out at this point that although the small-scale parameters characterizing the propagation channels linking the APs and MSs can only be safely assumed to be static over a coherence time-frequency interval of τ_c samples, the large-scale parameters (i.e., path loss propagation losses and spatial covariance matrices) can be safely assumed to be static over a time-frequency interval $\tau_{Lc} \gg \tau_c$ [12], [28]. These particular channel characteristics will be leveraged in the next subsections to simplify both the channel estimation and the precoding/combining processes.

A. SPATIAL MODELING OF THE MS DISTRIBUTION

As we are interested in exploring the impact ASO strategies may have on the performance of cell-free mmWave massive MIMO networks with a non-uniform spatial traffic distribution, the location of MSs on the service area will be modeled using the approach proposed by Lee *et al.* in [27]. This spatial traffic model generates large-scale spatial traffic variations by resorting to the use of sums of sinusoids capturing the characteristics of spatially correlated log-normally distributed traffic. In particular, let us consider an square area $\mathcal{S}$ of side D. This target region is tessellated in a regular grid of N_X by N_Y rectangular cells (or pixels). A cell (or pixel) (x, y) where $x \in \{1, \ldots, N_X\}$ and $y \in \{1, \ldots, N_Y\}$, is characterized by a traffic density demand $\rho_{x,y}$ (in MSs per pixel). To generate a log-normal distributed traffic map, a Gaussian random field is first produced as

$$\rho^G_{x,y} = \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \cos\left(i_t^{(\mathrm{u})} \Re\{p_{x,y}\} + \theta_t^{(\mathrm{u})}\right) \times \cos\left(j_t^{(\mathrm{u})} \Im\{p_{x,y}\} + \phi_t^{(\mathrm{u})}\right), \quad (1)$$

where $p_{x,y} = \Re\{p_{x,y}\} + j\Im\{p_{x,y}\}$ is used to denote the location (on a complex plane) of the center of pixel (x, y). The angular frequencies $i_t^{(\mathrm{u})}$ and $j_t^{(\mathrm{u})}$ are random variables uniformly distributed in the range $[0, \omega^{\mathcal{S}}_{\max}]$, where $\omega^{\mathcal{S}}_{\max}$ is defined as the maximum spatial spread used to control the rate of fluctuations of the random field in the area $\mathcal{S}$, and phases $\theta_t^{(\mathrm{u})}$ and $\phi_t^{(\mathrm{u})}$ are random variables uniformly distributed in the range $[0, 2\pi]$. According to the central limit theorem, for a large enough value of T, the symbol $\rho^G_{x,y}$ can be approximated as a standard Gaussian random variable. In [27], Lee *et al.* found that using $T = 10$ provides sufficiently accurate results.

By calculating the exponential function of $\rho^G_{x,y}$ with the location parameter $\mu^{\mathcal{S}}$ and the scaling parameter $\sigma^{\mathcal{S}}$, a spatial traffic density matrix can be obtained whose elements $\rho_{x,y}$, for all $x \in \{1, \ldots, N_X\}$ and $y \in \{1, \ldots, N_Y\}$, can be expressed as

$$\rho_{x,y} = \exp\left(\sigma^{\mathcal{S}} \rho^G_{x,y} + \mu^{\mathcal{S}}\right). \quad (2)$$

These random variables are log-normally distributed and by controlling the parameters $\mu^{\mathcal{S}}$ and $\sigma^{\mathcal{S}}$ the corresponding log-normal distribution can be scaled to fit the statistics of the traffic distribution experienced in different scenarios [27].

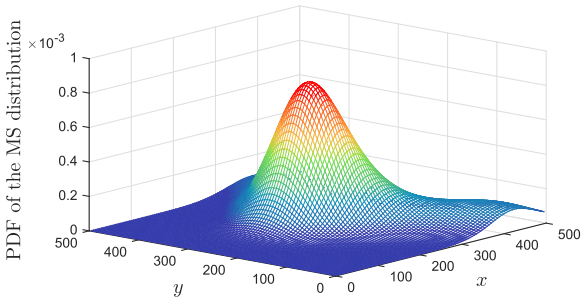


FIGURE 2. Probability density function of the number of MSs per pixel on a square grid of side $D = 500$ m and square pixels of side 5 m in an urban scenario (see parameters in [27, Table 1]).

The probability density function (pdf) of the number of MSs per pixel can be finally calculated as

$$f_{x,y}^{\mathrm{MS}} = \frac{\rho_{x,y}}{\sum_{x'=1}^{N_X}\sum_{y'=1}^{N_Y}\rho_{x',y'}}. \tag{3}$$

For instance, Fig. 2 represents the pdf characterizing the number of MSs per pixel on a square grid of side $D = 500$ m and square pixels of side 5 m using $\omega_{\max}^{\mathcal{S}}$ =0.012673, $\mu^{\mathcal{S}}$ =17.7956 and $\sigma^{\mathcal{S}}$ =2.1188, which are characteristic values of urban scenarios according to [27, Table 1].

B. CHANNEL MODEL

Compared to lower frequency bands, propagation in the mmWave band is characterized by very high distance- and penetration-based propagation losses that lead to sparse scattering multipath propagation, thus boosting the importance of line-of-sight (LOS) propagation, reflection and blockage. Furthermore, high antenna correlation levels may be expected because of the use of mmWave transmitters and receivers with tightly-packet large antenna arrays. The peculiarities of these mmWave channels can be captured by using a simplified clustered channel model version inspired by the third generation partnership project (3GPP) for Urban Micro-cell scenarios described in [29] and by a variety of research works (see [8], [16], [30], [31] and references therein).

According to this simplified clustered channel model, the propagation link between the mth AP and MS k can be either in outage, in LOS or in non-line-of-sight (NLOS) conditions with probabilities [8]

$$p_{\mathrm{out}}(d_{mk}) = \max\left(0, 1 - e^{-a_{\mathrm{out}}d_{mk}+b_{\mathrm{out}}}\right), \tag{4a}$$

$$p_{\mathrm{LOS}}(d_{mk}) = \left(1 - p_{\mathrm{out}}(d_{mk})\right) e^{-a_{\mathrm{LOS}}d_{mk}}, \tag{4b}$$

$$p_{\mathrm{NLOS}}(d_{mk}) = 1 - p_{\mathrm{out}}(d_{mk}) - p_{\mathrm{LOS}}(d_{mk}), \tag{4c}$$

respectively, where d_{mk} is the distance (in meters) between the AP and the MS. The parameters governing these probabilities are set to $1/a_{\mathrm{out}} = 30$ m, $b_{\mathrm{out}} = 5.2$, and $1/a_{\mathrm{LOS}} = 67.1$ m (see [8, Table 1]). When in outage conditions, this propagation link will be characterized by infinite propagation losses. When in LOS or NLOS conditions, however, a standard linear model with shadowing will be used that can be expressed as [29]

$$L_{mk}[dB] = \alpha + 10\beta\log_{10}(d_{mk}) + \chi_{mk}, \tag{5}$$

where α and β are frequency-dependent least square fits of floating intercept and slope, respectively, and are selected according to whether the link is in LOS or NLOS, and the large-scale shadow fading component χ_{mk} is modelled as a zero mean spatially correlated normal random variable with standard deviation σ_χ.

Based on the large-scale propagation loss model just described, the channel vector $\breve{\boldsymbol{h}}_{mk} \in \mathbb{C}^{N\times 1}$ from the kth MS to the mth AP (including both large-scale and small-scale fading) can be generically characterized as a Ricean fading channel consisting of a LOS component on top of a Rayleigh distributed component modelling the scattered multipath. That is,

$$\breve{\boldsymbol{h}}_{mk} = \sqrt{\frac{K_{mk}}{K_{mk}+1}}\overline{\boldsymbol{h}}_{mk} + \sqrt{\frac{1}{K_{mk}+1}}\boldsymbol{h}_{mk}, \tag{6}$$

with a normalized LOS component

$$\overline{\boldsymbol{h}}_{mk} = \alpha_{mk}\boldsymbol{a}\left(\overline{\theta}_{mk,1}, \overline{\phi}_{mk,1}\right), \tag{7}$$

and a normalized NLOS component

$$\boldsymbol{h}_{mk} = \sum_{c=1}^{C_{mk}}\sum_{p=1}^{P_{mk}}\alpha_{mk,cp}\boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right), \tag{8}$$

where K_{mk} is the Ricean K-factor, with $K_{mk} = 0$ for NLOS propagation links and $10\log_{10}(K_{mk}) \sim \mathcal{N}\left(\mu_K, \sigma_K^2\right)$ for LOS propagation links. The parameter $\alpha_{mk} = 10^{-L_{mk}/20}e^{j\kappa_{mk}}$, with $\kappa_{mk} \sim \mathcal{U}[0, 2\pi]$, is used to denote the large-scale complex channel gain of the LOS component, C_{mk} and P_{mk} are the number of contributing scattering clusters of the NLOS component and the number of propagation paths per cluster, respectively, $\alpha_{mk,cp}$ is the complex small-scale fading gain on the pth path of cluster c, and $\boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right)$ is the normalized array response vector of the AP at the azimuth and elevation angles $\theta_{mk,cp}$ and $\phi_{mk,cp}$, respectively.

As suggested by Akdeniz *et al.* in [8], $\theta_{mk,cp}$ can be generated as a wrapped Gaussian around the cluster central angle $\overline{\theta}_{mk,c}$ with standard deviation given by the root mean square (rms) azimuth angular spreads for the cluster. Furthermore, $\phi_{mk,cp}$ can be generated as a Laplacian around the cluster central angle $\overline{\phi}_{mk,c}$ with scale parameters given by the rms elevation angular spreads for the cluster. The azimuth cluster central angle $\overline{\theta}_{mk,c}$ is uniformly distributed in the range $[-\pi, \pi]$ and the elevation cluster central angle $\overline{\phi}_{mk,c}$ is set to the corresponding LOS elevation angle. The cluster rms angular spreads are exponentially distributed with a mean equal to $1/\lambda_{\mathrm{rms}}$ that depends on whether we are considering the azimuth or elevation directions. The small-scale scattering fading gains are distributed as

$$\alpha_{mk,cp} \sim \mathcal{CN}\left(0, \gamma_{mk,c}10^{-L_{mk}/10}\right), \tag{9}$$

where the cluster c is assumed to contribute to the scatter fading with a fraction of power given by

$$\gamma_{mk,c} = \frac{N\gamma'_{mk,c}}{P_{mk}\sum_{j=1}^{C_{mk}}\gamma'_{mk,j}}, \tag{10}$$

with

$$\gamma'_{mk,j} = U_{mk,j}^{r_\tau - 1} 10^{Z_{mk,j}/10}, \tag{11}$$

where $U_{mk,j} \sim \mathcal{U}[0, 1]$, $Z_{mk,j} \sim \mathcal{N}(0, \zeta^2)$, and the constants r_τ and ζ^2 being treated as model parameters [8].

Using this channel propagation model, the spatial covariance matrix of the scattered multipath component $\boldsymbol{h}_{mk}$ can be obtained as

$$\begin{aligned}\boldsymbol{R}_{mk} &= \mathbb{E}\left\{\boldsymbol{h}_{mk}\boldsymbol{h}_{mk}^H\right\} \\ &= 10^{-L_{mk}/10} \sum_{c=1}^{C_{mk}} \gamma_{mk,c} \\ &\quad \times \sum_{p=1}^{P_{mk}} \boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right)\left(\boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right)\right)^H. \end{aligned} \tag{12}$$

and the spatial covariance matrix of the resulting channel vector $\check{\boldsymbol{h}}_{mk}$ can thus be expressed as

$$\begin{aligned}\check{\boldsymbol{R}}_{mk} &= \mathbb{E}\left\{\check{\boldsymbol{h}}_{mk}\check{\boldsymbol{h}}_{mk}^H\right\} \\ &= \frac{K_{mk}}{K_{mk}+1}\bar{\boldsymbol{h}}_{mk}\bar{\boldsymbol{h}}_{mk}^H + \frac{\boldsymbol{R}_{mk}}{K_{mk}+1}. \end{aligned} \tag{13}$$

As stated by Özdogan *et al.* in [32], these spatial covariance matrices, as well as the propagation path losses, Ricean K-factors, and channel means, can be considered to be constant over frequency-time intervals much larger than the coherence interval τ_c and, consequently, they can be straightforwardly estimated in practice using the sample mean and sample covariance matrices [33]–[36].

C. RF PRECODER/COMBINER DESIGN

Without loss of essential generality, it is assumed in this paper that a hybrid precoding technique is used in which each RF chain is dedicated to one and only one MS. In particular, if $K \leq L$, all active APs in the network provide service to the K MSs and only K RF chains per AP are activated (one for each MS in the network). If $K > L$, instead, each active AP can only serve a subset of L MSs (one for each RF chain) and thus, an algorithm must be devised to decide which are the subsets of MSs to be beamformed from each of the active APs while ensuring that the K MSs are simultaneously served. The number of active RF chains per AP can then be generically expressed as $L_A = \min\{K, L\}$ (see Fig. 1b).

Channel reciprocity is exploited by implementing an $N \times L_A$ RF precoding matrix $\boldsymbol{W}_m^{\mathrm{RF}}$, describing the effects of the active analog phase shifters at the mth active AP, which is common to the UL (RF combining phase) and DL (RF precoding phase). Furthermore, denoting by $\mathcal{K}_m = \{\kappa_{m1}, \ldots, \kappa_{mL_A}\}$ the set of L_A MSs beamformed by the mth AP, it is assumed that $\boldsymbol{W}_m^{\mathrm{RF}}$ is a function of only the spatial channel covariance matrices $\left\{\check{\boldsymbol{R}}_{mk}\right\}_{k\in\mathcal{K}_m}$, known at the mth AP through spatial channel covariance estimation for hybrid analog-digital MIMO precoding architectures [37]–[39]. The use of long-term channel statistics such as the spatial covariance matrices is a reasonable approach as they vary over very long time scales and, moreover, they can be safely assumed to be uniform across the whole system bandwidth, thus providing a good solution to the problem of designing a common analog precoder for all subcarriers [39].

Algorithm 1 Selection of MSs to Beamform From Each AP

Input: $\xi_{mk}\ \forall mk$, $\mathcal{M}^A$, M_A, L, K
Initialization: $\mathcal{K}_m^{(0)} = \{1, \ldots, K\}\ \forall m$
$\mathcal{M}_k^{(0)} = \mathcal{M}^A\ \forall k$
$\mathcal{A} = \mathcal{M}^A$
for $i = 1 : M_A(K - L)$ **do**
{Find edge producing max-min $\xi_k^{(i)}$ when removed}
$(m^*, k^*) = \arg \max_{k\in\{1,\ldots,K\}} \min_{m\in\mathcal{A}} \sum_{n\in\mathcal{M}_k^{(i-1)}\setminus m} \xi_{nk}$
{Remove k^* from the set of MSs served by AP m^*}
$\mathcal{K}_{m^*}^{(i)} = \mathcal{K}_{m^*}^{(i-1)} \setminus k^*$
{Remove m^* from the sets of APs serving MS k^*}
$\mathcal{M}_{k^*}^{(i)} = \mathcal{M}_{k^*}^{(i-1)} \setminus m^*$
{Update set of APs already serving L MSs}
if $\left|\mathcal{K}_{m^*}^{(i)}\right| = L$ **then**
$\mathcal{A} = \mathcal{A} \setminus m^*$
end if
end for
Output: $\mathcal{K}_m = \mathcal{K}_m^{(M_A(K-L))}\ \forall m \in \mathcal{M}^A$

The Hermitian covariance matrix of the propagation channel linking MS k and AP m can be factorized using eigendecomposition as $\check{\boldsymbol{R}}_{mk} = \boldsymbol{U}_{mk}\boldsymbol{\Lambda}_{mk}\boldsymbol{U}_{mk}^H$, where $\boldsymbol{\Lambda}_{mk} = \mathrm{diag}\left(\left[\lambda_{mk,1}\ \ldots\ \lambda_{mk,r_{mk}}\right]\right)$ contains the r_{mk} non-null eigenvalues of $\check{\boldsymbol{R}}_{mk}$, and $\boldsymbol{U}_{mk}$ is the $N \times r_{mk}$ matrix of the corresponding eigenvectors. This eigen-factorization can be exploited to design the analog RF precoder/combiner stage by using the well-known (constrained) statistical eigen beamforming [39], [40], where

$$\begin{aligned}\boldsymbol{W}_m^{\mathrm{RF}} &= \left[\boldsymbol{w}_{m\kappa_{m1}}^{\mathrm{RF}}\ \ldots\ \boldsymbol{w}_{m\kappa_{mL_A}}^{\mathrm{RF}}\right] \\ &= \left[e^{-j\angle \boldsymbol{u}_{m\kappa_{m1},\max}}\ \ldots\ e^{-j\angle \boldsymbol{u}_{m\kappa_{mL_A},\max}}\right], \end{aligned} \tag{14}$$

with $\boldsymbol{u}_{mk,\max}$ denoting the dominant eigenvector of $\check{\boldsymbol{R}}_{mk}$ associated to the maximum eigenvalue $\lambda_{mk,\max}$, and the function $\angle\boldsymbol{x}$ returning the phase angles, in radians, for each element of the complex vector $\boldsymbol{x}$. The equivalent channel vector between MS k and AP m, including the analog RF precoder/combiner, can then be defined as

$$\boldsymbol{g}_{mk} = \boldsymbol{W}_m^{\mathrm{RF}^T}\check{\boldsymbol{h}}_{mk} \in \mathbb{C}^{L_A\times 1}, \tag{15}$$

whose dimension L_A is typically much less than the number of antennas of the massive MIMO array, thus making the small-scale training phase computationally simpler.

D. SELECTION OF MSs TO BEAMFORM FROM EACH AP

As we only consider the use of analog precoding/decoding stages in which each RF chain at a given AP is dedicated

to a single MS, in those cases in which the number of MSs is greater than the number of available RF chains at each AP (i.e., $K > L$), the group of L MSs to beamform from each active AP in the cell-free network, indexed by the sets $\mathcal{K}_m = \{\kappa_{m1}, \ldots, \kappa_{mL}\}$, for all $m \in \{1, \ldots, M\}$, will have to be selected. Furthermore, as the RF beamforming/decoding matrices at the APs are designed assuming only the availability of the large-scale spatial channel covariance matrices, this selection process can only be based on this large-scale CSI. Inspired by the Frobenius norm-based suboptimal user selection algorithm proposed by Shen *et al.* in [41], an iterative selection algorithm was proposed in [16] that, under the constraint that each AP can only beamform to L MSs, aims at maximizing the minimum average sum energy of the equivalent channels between the M_A active APs and any of the K MSs in the network. Note that, using this algorithm, each active AP will beamform to exactly L MSs and each MS will be beamformed by at least one AP.

At the beginning of the ith iteration of the algorithm, a simple edge-weighted directed graph with M_A source nodes and K sink nodes is used to represent the cell-free massive MIMO network. In this directed graph, the mth source node, which represents the mth active AP, is connected to a group $\mathcal{K}_m^{(i)}$ of sink nodes, used to represent the MSs to be *potentially* beamformed from the mth active AP. The average energy of the equivalent channel linking the mth active AP and MS $l \in \mathcal{K}_m^{(i)}$, which can be obtained as

$$\xi_{ml} = \mathbb{E}\left\{\left|\boldsymbol{w}_{ml}^{\mathrm{RF}^T}\check{\boldsymbol{h}}_{ml}\right|^2\right\} = \boldsymbol{w}_{ml}^{\mathrm{RF}^T}\check{\boldsymbol{R}}_{ml}\boldsymbol{w}_{ml}^{\mathrm{RF}^*}, \tag{16}$$

is used to weight the connection (edge) joining the mth source node and the lth sink node in $\mathcal{K}_m^{(i)}$. Using this notation, the average sum energy of the equivalent channels between the M_A active APs and MS k at the beginning of the ith iteration can be obtained as

$$\mathcal{E}_k^{(i)} = \sum_{m \in \mathcal{M}_k^{(i)}} \xi_{mk}, \tag{17}$$

where $\mathcal{M}_k^{(i)}$ is the set of active APs selected in previous iterations to beamform to MS k. The reverse-delete algorithm is used in this iteration to remove the edge (i.e., the RF chain and associated beamformer) coming from one of those active APs still beamforming to more than L MSs that maximizes the minimum average sum energy per MS after removal. The proposed algorithm, starting with a fully connected graph, stops after $M_A(K - L)$ iterations with all active APs in $\mathcal{M}^A$ beamforming to exactly L MSs. A mathematical pseudocode for this algorithm is shown in Algorithm 1.

E. SMALL-SCALE TRAINING PHASE: CHANNEL ESTIMATION

During the UL training phase, all K MSs simultaneously transmit pilot sequences of τ_p samples to the active APs and thus, the $L_A \times \tau_p$ received UL signal matrix at the mth active AP can be expressed as

$$\boldsymbol{Y}_{p_m} = \sqrt{\tau_p P_p}\sum_{k'=1}^{K}\boldsymbol{g}_{mk'}\boldsymbol{\varphi}_{k'}^T + \boldsymbol{N}_{p_m}, \tag{18}$$

where P_p is the transmit power of each pilot symbol, $\boldsymbol{\varphi}_k$ denotes the $\tau_p \times 1$ pilot signal allocated to MS k, with $\|\boldsymbol{\varphi}_k\|^2 = 1$, and $\boldsymbol{N}_{p_m}$ is an $L_A \times \tau_p$ matrix of i.i.d. additive noise samples with each entry distributed as[1] $\mathcal{CN}(0, \sigma_u^2(N))$. Note that, since in most practical scenarios it holds that $K > \tau_p$, a given pilot sequence will be allocated to more than one MS and, hence, pilot contamination will arise [28], [42].

As previously stated, considering scenarios where MSs move slowly, it is reasonable to assume that the Ricean K-factors K_{mk}, the LOS components $\overline{\boldsymbol{h}}_{mk}$, and the scatter fading correlation matrices $\boldsymbol{R}_{mk}$ change slowly and can be perfectly known at the mth active AP, for all k [43]. Under this assumption, we can define

$$\begin{aligned}\check{\boldsymbol{y}}_{p\,mk} &= \left(\boldsymbol{Y}_{p_m} - \mathbb{E}\left\{\boldsymbol{Y}_{p_m}\right\}\right)\boldsymbol{\varphi}_k^* \\ &= \left(\sum_{k'=1}^{K}\sqrt{\frac{\tau_p P_p}{K_{mk'}+1}}\boldsymbol{W}_m^{\mathrm{RF}^T}\boldsymbol{h}_{mk'}\boldsymbol{\varphi}_{k'}^T + \boldsymbol{N}_{p_m}\right)\boldsymbol{\varphi}_k^*\end{aligned} \tag{19}$$

and

$$\check{\boldsymbol{g}}_{mk} = \boldsymbol{g}_{mk} - \mathbb{E}\left\{\boldsymbol{g}_{mk}\right\} = \sqrt{\frac{1}{K_{mk}+1}}\boldsymbol{W}_m^{\mathrm{RF}^T}\boldsymbol{h}_{mk}, \tag{20}$$

and then derive the minimum mean square error (MMSE) estimate for the channel between the kth MS and the mth active AP as [43], [44]

$$\begin{aligned}\hat{\boldsymbol{g}}_{mk} &= \sqrt{\frac{K_{mk}}{K_{mk}+1}}\boldsymbol{W}_m^{\mathrm{RF}^T}\overline{\boldsymbol{h}}_{mk} \\ &\quad + \mathbb{E}\left\{\check{\boldsymbol{y}}_{p\,mk}\check{\boldsymbol{g}}_{mk}^H\right\}\left(\mathbb{E}\left\{\check{\boldsymbol{y}}_{p\,mk}\check{\boldsymbol{y}}_{p\,mk}^H\right\}\right)^{-1}\check{\boldsymbol{y}}_{p\,mk} \\ &= \sqrt{\frac{K_{mk}}{K_{mk}+1}}\boldsymbol{W}_m^{\mathrm{RF}^T}\overline{\boldsymbol{h}}_{mk} \\ &\quad + \frac{\sqrt{\tau_p P_p}}{K_{mk}+1}\boldsymbol{R}_{mk}^{\mathrm{RF}}\boldsymbol{\Psi}_{mk}^{-1}\check{\boldsymbol{y}}_{p\,mk},\end{aligned} \tag{21}$$

where

$$\boldsymbol{R}_{mk}^{\mathrm{RF}} = \boldsymbol{W}_m^{\mathrm{RF}^T}\boldsymbol{R}_{mk}\boldsymbol{W}_m^{\mathrm{RF}^*}, \tag{22}$$

[1] Note that in the UL of a fully-connected hybrid beamforming architecture each reception chain is composed of N antenna elements, each connected to a low-noise amplifier (LNA) characterized by a power gain G_{LNA} and a noise temperature T_{LNA}. Each of the N LNAs feeds an analog passive phase shifter characterized by an insertion loss L_{PS}. The outputs of the N phase shifters are introduced to a power combiner whose insertion losses are typically proportional to the number of inputs, that is, $L_{\mathrm{PC}} = NL_{\mathrm{PC}_{in}}$. Finally, the output of the power combiner is introduced to an RF chain characterized by a power gain G_{RF} and a noise temperature T_{RF}. Thus, the equivalent noise temperature of each receive chain can be obtained as $T_u = N\left(T_0 + T_{\mathrm{LNA}} + \frac{T_0(L_{\mathrm{PS}}L_{\mathrm{PC}_{in}}-1)}{G_{\mathrm{LNA}}} + \frac{T_{\mathrm{RF}}L_{\mathrm{PS}}L_{\mathrm{PC}_{in}}}{G_{\mathrm{LNA}}}\right)$.

and

$$\boldsymbol{\Psi}_{mk} = \tau_p P_p \sum_{k'=1}^{K} \frac{\boldsymbol{R}_{mk'}^{\mathrm{RF}}}{K_{mk'}+1} \left|\boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k\right|^2 + \sigma_u^2(N)\boldsymbol{I}_{L_A}. \quad (23)$$

The channel estimate $\hat{\boldsymbol{g}}_{mk}$ and the MMSE channel estimation error $\tilde{\boldsymbol{g}}_{mk} = \boldsymbol{g}_{mk} - \hat{\boldsymbol{g}}_{mk}$ are uncorrelated random vectors distributed as

$$\hat{\boldsymbol{g}}_{mk} \sim \mathcal{CN}\left(\sqrt{\frac{K_{mk}}{K_{mk}+1}}\boldsymbol{W}_m^{\mathrm{RF}\,T}\overline{\boldsymbol{h}}_{mk}, \hat{\boldsymbol{A}}_{mk}\right), \quad (24)$$

and $\tilde{\boldsymbol{g}}_{mk} \sim \mathcal{CN}\left(\boldsymbol{0}, \tilde{\boldsymbol{A}}_{mk}\right)$, respectively, where

$$\hat{\boldsymbol{A}}_{mk} = \frac{\tau_p P_p \boldsymbol{R}_{mk}^{\mathrm{RF}} \boldsymbol{\Psi}_{mk}^{-1} \left(\boldsymbol{R}_{mk}^{\mathrm{RF}}\right)^H}{(K_{mk}+1)^2}, \quad (25)$$

is the covariance matrix of $\hat{\boldsymbol{g}}_{mk}$ and

$$\tilde{\boldsymbol{A}}_{mk} = \mathbb{E}\left\{\tilde{\boldsymbol{g}}_{mk}\tilde{\boldsymbol{g}}_{mk}^H\right\} = \frac{\boldsymbol{R}_{mk}^{\mathrm{RF}}}{K_{mk}+1} - \hat{\boldsymbol{A}}_{mk} \quad (26)$$

is the covariance matrix of $\tilde{\boldsymbol{g}}_{mk}$.

F. DOWNLINK PAYLOAD DATA TRANSMISSION

Let us define $\boldsymbol{s}_d = [s_{d1} \ldots s_{dK}]^T$ as the $K \times 1$ vector of symbols jointly transmitted to the K MSs, such that $E\left\{\boldsymbol{s}_d \boldsymbol{s}_d^H\right\} = \boldsymbol{I}_K$. Assuming the use of a centralized baseband precoder at the CPU, symbol vector $\boldsymbol{s}_d$ undergoes some signal processing operations before being transmitted, including a power allocation process and a baseband precoding task at the CPU, and an RF precoding process at the APs. Thus, the transmitted signal vector from the mth active AP can be generically expressed as

$$\boldsymbol{x}_m = \boldsymbol{W}_m^{\mathrm{RF}} \boldsymbol{W}_{d\,m}^{\mathrm{BB}} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d, \quad (27)$$

with $\boldsymbol{W}_{d\,m}^{\mathrm{BB}} = \left[\boldsymbol{w}_{dm1}^{\mathrm{BB}} \ldots \boldsymbol{w}_{dmK}^{\mathrm{BB}}\right] \in \mathbb{C}^{L_A \times K}$ denoting the baseband precoding matrix affecting the signal transmitted by the mth active AP, and $\boldsymbol{\Upsilon} = \mathrm{diag}\left([\upsilon_1 \ldots \upsilon_K]\right)$ being a $K \times K$ diagonal matrix containing the power control coefficients in its main diagonal. The power control coefficients are chosen to satisfy the power constraints

$$\mathbb{E}\left\{\|\boldsymbol{x}_m\|^2\right\} = \sum_{k=1}^{K} \upsilon_k \theta_{mk}^{\mathrm{BB/RF}} \leq \overline{P}_m, \quad (28)$$

for all $m \in \mathcal{M}^A$, where we have used the definition

$$\theta_{mk}^{\mathrm{BB/RF}} = \mathbb{E}\left\{\left\|\boldsymbol{W}_m^{\mathrm{RF}} \boldsymbol{w}_{dmk}^{\mathrm{BB}}\right\|^2\right\}, \quad (29)$$

and $\overline{P}_m$ is the maximum average transmit power available at AP m. Using this notation, the signal received by MS k can be expressed as

$$y_{dk} = \sum_{m \in \mathcal{M}^A} \breve{\boldsymbol{h}}_{mk}^T \boldsymbol{x}_m + n_{dk}, \quad (30)$$

where $n_{dk} \sim \mathcal{CN}(0, \sigma_d^2)$ is the noise sample at MS k.

The vector $\boldsymbol{y}_d = \left[y_{d1} \ldots y_{dK}\right]^T$ containing the signals received by the K scheduled MSs in the network can be written as

$$\begin{aligned}\boldsymbol{y}_d &= \sum_{m \in \mathcal{M}^A} \breve{\boldsymbol{H}}_m^T \boldsymbol{x}_m + \boldsymbol{n}_d \\ &= \sum_{m \in \mathcal{M}^A} \breve{\boldsymbol{H}}_m^T \boldsymbol{W}_m^{\mathrm{RF}} \boldsymbol{W}_{d\,m}^{\mathrm{BB}} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d + \boldsymbol{n}_d \\ &= \boldsymbol{G}^T \boldsymbol{W}_d^{\mathrm{BB}} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d + \boldsymbol{n}_d, \end{aligned} \quad (31)$$

where $\breve{\boldsymbol{H}}_m = \left[\breve{\boldsymbol{h}}_{m1} \ldots \breve{\boldsymbol{h}}_{mK}\right] \in \mathbb{C}^{L \times K}$ represents the MIMO channel between the mth active AP and the K MSs, $\boldsymbol{W}_d^{\mathrm{BB}} = \left[\boldsymbol{W}_{d\,m_1^A}^{\mathrm{BB}\,T} \ldots \boldsymbol{W}_{d\,m_{M_A}^A}^{\mathrm{BB}\,T}\right]^T \in \mathbb{C}^{M_A L_A \times K}$ is the digital precoding filter stage implemented at the CPU, $\boldsymbol{G} = [\boldsymbol{G}_{m_1^A}^T \ldots \boldsymbol{G}_{m_{M_A}^A}^T]^T \in \mathbb{C}^{M_A L_A \times K}$ is the global equivalent MIMO channel (including the RF precoding/decoding matrices) between the K MSs and the digital processing stage at the CPU, with $\boldsymbol{G}_m = \boldsymbol{W}_m^{\mathrm{RF}\,T} \breve{\boldsymbol{H}}_m$ representing the equivalent MIMO channel matrix between the K MSs and the digital processing stage corresponding to the mth active AP and, finally, $\boldsymbol{n}_d = [n_{d1} \ldots n_{dK}]^T$ is the vector containing the noise samples at the MSs.

Assuming the use of the classical ZF multiuser-MIMO (MU-MIMO) baseband precoder to tackle the spatial multiplexing, we have that

$$\boldsymbol{W}_d^{\mathrm{BB}} = \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \quad (32)$$

or, equivalently,

$$\boldsymbol{W}_{d\,m}^{\mathrm{BB}} = \hat{\boldsymbol{G}}_m^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \ \forall m \in \mathcal{M}^A, \quad (33)$$

where we have assumed that $\boldsymbol{G} = \hat{\boldsymbol{G}} + \tilde{\boldsymbol{G}}$ and $\boldsymbol{G}_m = \hat{\boldsymbol{G}}_m + \tilde{\boldsymbol{G}}_m$. Consequently, the signal received by the kth MS can be expressed as

$$\begin{aligned} y_{dk} &= \boldsymbol{g}_k^T \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d + n_{dk} \\ &= \left(\hat{\boldsymbol{g}}_k^T + \tilde{\boldsymbol{g}}_k^T\right) \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d + n_{dk} \\ &= \sqrt{\upsilon_k} s_{dk} + \tilde{\boldsymbol{g}}_k^T \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}_d + n_{dk} \end{aligned} \quad (34)$$

The first term denotes the useful received signal, the second term contains the interference terms due to the use of imperfect CSI (pilot contamination), and the third term is the thermal noise sample.

G. UPLINK PAYLOAD DATA TRANSMISSION

In the UL, the vector of received signals at the output of the L_A RF chains (including the RF phase shifters) of the mth active AP is given by

$$\begin{aligned}\boldsymbol{r}_{um} &= \sqrt{P_u} \sum_{k'=1}^{K} \boldsymbol{g}_{mk'} \sqrt{\omega_{k'}} s_{uk'} + \boldsymbol{n}_{um} \\ &= \sqrt{P_u} \boldsymbol{G}_m \boldsymbol{\Omega}^{1/2} \boldsymbol{s}_u + \boldsymbol{n}_{um}, \end{aligned} \quad (35)$$

where P_u is the maximum average UL transmit power available at any of the active MSs, $\boldsymbol{s}_u = [s_{u1} \ldots s_{uK}]^T$ denotes the vector of symbols transmitted by the K MS, $\boldsymbol{\Omega} = \text{diag}([\omega_1 \ldots \omega_K])$, with $0 \leq \omega_k \leq 1$, is a matrix containing the power control coefficients used at the MSs, and $\boldsymbol{n}_{um} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_u^2(N)\boldsymbol{I}_{L_A})$ is the vector of additive thermal noise samples at the output of the L_A RF chains of the mth active AP. The received vector of signals at each of the active APs in $\mathcal{M}^A$ is forwarded to the CPU where it is processed using a baseband combining matrix. In particular, assuming the use of ZF MIMO detection, the CPU uses the detection matrix

$$\boldsymbol{W}_u^{\text{BB}} = \left(\hat{\boldsymbol{G}}^H \hat{\boldsymbol{G}}\right)^{-1} \hat{\boldsymbol{G}}^H = \boldsymbol{W}_d^{\text{BB}\,T} \tag{36}$$

or, equivalently

$$\boldsymbol{W}_{um}^{\text{BB}} = \left(\hat{\boldsymbol{G}}^H \hat{\boldsymbol{G}}\right)^{-1} \hat{\boldsymbol{G}}_m^H = \boldsymbol{W}_{dm}^{\text{BB}\,T}, \quad \forall m \in \mathcal{M}^A, \tag{37}$$

to jointly process the vector $\boldsymbol{z}_u = \left[\boldsymbol{z}_{u1}^T \ldots \boldsymbol{z}_{uM}^T\right]^T$ and obtain the vector of detected samples

$$\begin{aligned} \boldsymbol{y}_u &= \boldsymbol{W}_u^{\text{BB}} \boldsymbol{z}_u = \sqrt{P_u} \boldsymbol{W}_u^{\text{BB}} \boldsymbol{G} \boldsymbol{\Omega}^{1/2} \boldsymbol{s}_u + \boldsymbol{\eta}_u \\ &= \sqrt{P_u} \boldsymbol{\Omega}^{1/2} \boldsymbol{s}_u + \sqrt{P_u} \boldsymbol{W}_u^{\text{BB}} \tilde{\boldsymbol{G}} \boldsymbol{\Omega}^{1/2} \boldsymbol{s}_u + \boldsymbol{\eta}_u, \end{aligned} \tag{38}$$

where $\boldsymbol{\eta}_u = \boldsymbol{W}_u^{\text{BB}} \boldsymbol{n}_u$. Again, the first term denotes the useful received signal, the second term contains the interference terms due to the use of imperfect CSI, and the third term includes the thermal noise samples. The detected sample corresponding to the symbol transmitted by the kth MS can then be obtained as

$$y_{uk} = \sqrt{P_u} \omega_k^{1/2} s_{uk} + \sqrt{P_u} \left[\boldsymbol{W}_u^{\text{BB}} \tilde{\boldsymbol{G}} \boldsymbol{\Omega}^{1/2} \boldsymbol{s}_u\right]_k + \eta_{uk}, \tag{39}$$

where $[\boldsymbol{x}]_k$ denotes the kth entry of vector $\boldsymbol{x}$.

III. PERFORMANCE METRICS

A. SPECTRAL EFFICIENCY

Analysis techniques similar to those applied, for instance, in [3], [5], [28], [45]–[47], are used in this section to derive DL and UL spectral efficiencies (also known as achievable rates). In particular, the sum of the second and third terms on the right hand side (RHS) of (34), for the DL case, and (39), for the UL case, are treated as *effective noise*. The additive terms constituting the *effective noise* are, in both DL and UL cases, mutually uncorrelated, and uncorrelated with s_{dk} and s_{uk}, respectively. Therefore, both the desired signal and the so-called *effective noise* are uncorrelated. Now, recalling the fact that uncorrelated Gaussian noise represents the worst case, from a capacity point of view, and that the complex-valued fast fading random variables characterizing the propagation channels between different pairs of AP-MS connections are independent, the DL and UL spectral efficiencies (measured in bits per second per Hertz) can be obtained as follows. The DL spectral efficiency is given by

$$S_{ed}(\boldsymbol{\upsilon}) = \sum_{k=1}^{K} S_{edk}(\boldsymbol{\upsilon}) = \frac{\tau_d}{\tau_c} \sum_{k=1}^{K} \log_2\left(1 + \text{SINR}_{dk}\right), \tag{40}$$

with

$$\text{SINR}_{dk} = \frac{\upsilon_k}{\sum_{k'=1}^{K} \upsilon_{k'} \varpi_{kk'} + \sigma_d^2}, \tag{41}$$

where

$$\varpi_{kk'} = \left[\text{diag}\left(\mathbb{E}\left\{\boldsymbol{W}_d^{\text{BB}\,H} \tilde{\boldsymbol{g}}_k^* \tilde{\boldsymbol{g}}_k^T \boldsymbol{W}_d^{\text{BB}}\right\}\right)\right]_{k'}. \tag{42}$$

Analogously, the UL spectral efficiency is given by

$$S_{eu}(\boldsymbol{\omega}) = \sum_{k=1}^{K} S_{euk}(\boldsymbol{\omega}) = \frac{\tau_u}{\tau_c} \sum_{k=1}^{K} \log_2\left(1 + \text{SINR}_{uk}\right), \tag{43}$$

with

$$\text{SINR}_{uk} = \frac{P_u \omega_k}{P_u \sum_{k'=1}^{K} \omega_{k'} \delta_{kk'} + \sigma_{\eta_{uk}}^2(N)}, \tag{44}$$

where

$$\delta_{kk'} = \left[\text{diag}\left(\mathbb{E}\left\{\boldsymbol{W}_u^{\text{BB}} \tilde{\boldsymbol{g}}_{k'} \tilde{\boldsymbol{g}}_{k'}^H \boldsymbol{W}_u^{\text{BB}\,H}\right\}\right)\right]_k, \tag{45}$$

and

$$\sigma_{\eta_{uk}}^2(N) = \sigma_u^2(N) \left[\text{diag}\left(\mathbb{E}\left\{\boldsymbol{W}_u^{\text{BB}} \boldsymbol{W}_u^{\text{BB}\,H}\right\}\right)\right]_k. \tag{46}$$

B. POWER CONSUMPTION MODEL

In a cell-free massive MIMO network implementing an ASO strategy, APs can be either in *active* or *sleep* mode. Moreover, an AP in active mode can be either receiving signals during the UL training and payload data transmission phases, or transmitting information during the DL payload data transmission phase. When in active mode, the power consumption at the mth AP depends on the UL spectral efficiency $S_{eu}(\boldsymbol{\omega})$ during the UL payload data transmission phase or on the radiated power P_m^{tx} during the DL payload data transmission phase. But it also depends on parameters such as the efficiency of the power amplifier, the power consumed by the small-signal RF transceiver and the baseband circuitry, or the losses produced by the feeder, the DC-DC power supply, the main supply, or the cooling system [48]–[52]. When in the sleep mode, the AP is kept in a state allowing a fairly rapid activation and hence it is not completely turned off. It is thus in a reduced power consumption state in which, although it is not radiating or receiving power, there are components such as the power supply, some of the signal processing blocks, and part of the cooling system that are still active and thus consuming power. Accordingly, a linear model can be used to approximate the total power consumed at the mth AP as (see, for instance, [48]–[52] and references therein)

$$P_m^{\text{AP}} = \begin{cases} \dfrac{P_m^{\text{tx}}(\boldsymbol{\upsilon})}{\alpha_m^{\text{AP}}} + P_{m\,d}^{\text{AP,fix}} + L_A P_{m\,d}^{\text{AP,chain}} & \text{DL Active} \\ B\xi_m^{\text{AP}} S_{eu}(\boldsymbol{\omega}) + P_{m\,u}^{\text{AP,fix}} + L_A P_{m\,u}^{\text{AP,chain}} & \text{UL Active} \\ P_{m\,\text{sleep}}^{\text{AP,fix}} + L_A P_{m\,\text{sleep}}^{\text{AP,chain}} & \text{Sleep,} \end{cases} \tag{47}$$

where α_m^{AP} is the power amplifier efficiency at the mth AP, B is the system bandwidth, ξ_m^{AP} is the traffic-dependent power consumption coefficient (in Watt per bit/s), $P_{m\,d}^{\mathrm{AP,fix}}$ and $P_{m\,u}^{\mathrm{AP,fix}}$ denote, respectively, the DL and UL power consumption figures that are independent of both the number of RF chains and the traffic load, $P_{m\,d}^{\mathrm{AP,chain}}$ and $P_{m\,u}^{\mathrm{AP,chain}}$ model the DL and UL traffic-independent power consumed by the circuitry related to each RF chain of the mth AP, respectively and, finally, $P_{m\,\mathrm{sleep}}^{\mathrm{AP,fix}}$ and $P_{m\,\mathrm{sleep}}^{\mathrm{AP,chain}}$ are the RF chain-independent and RF chain-dependent power consumed by the mth AP when in sleep mode.

A similar power consumption model can be established for the fronthaul links connecting the APs to the CPU. In particular, the power consumed by the mth fronthaul link when in active mode depends on the amount of traffic it has to convey and, thus, the total power consumption can be approximated as [2], [53]

$$P_m^{\mathrm{FH}} = \begin{cases} B\xi_m^{\mathrm{FH}} S_{ed}(\boldsymbol{\upsilon}) + P_m^{\mathrm{FH,fix}} & \text{DL Active} \\ B\xi_m^{\mathrm{FH}} S_{eu}(\boldsymbol{\omega}) + P_m^{\mathrm{FH,fix}} & \text{UL Active} \\ P_{m\,\mathrm{sleep}}^{\mathrm{FH,fix}} & \text{Sleep,} \end{cases} \tag{48}$$

where ξ_m^{FH} is the traffic-dependent power consumption coefficient (in Watt per bit/s), $P_m^{\mathrm{FH,fix}}$ is the traffic-independent power consumption when in active mode, and $P_{m\,\mathrm{sleep}}^{\mathrm{FH,fix}}$ accounts for the power consumed by the mth fronthaul link when in sleep mode.

The power consumption model for the MSs can also be approximated as

$$P_k^{\mathrm{MS}} = \begin{cases} B\xi_k^{\mathrm{MS}} S_{edk}(\boldsymbol{\upsilon}) + P_{k\,d}^{\mathrm{MS,fix}} & \text{DL} \\ \dfrac{P_u \omega_k}{\alpha_k^{\mathrm{MS}}} + P_{k\,u}^{\mathrm{MS,fix}} & \text{UL,} \end{cases} \tag{49}$$

where, again, α_k^{MS} is the power amplifier efficiency at the kth MS, ξ_k^{MS} is the traffic-dependent power consumption coefficient (in Watt per bit/s), $P_{k\,d}^{\mathrm{MS,fix}}$ and $P_{k\,u}^{\mathrm{MS,fix}}$ model the power consumed by the internal circuitry of the MS independently of the average radiated power, and $S_{edk}(\boldsymbol{\upsilon})$ denotes the DL spectral efficiency of the kth MS.

Putting all the pieces together, the total power consumption of the cell-free massive-MIMO network can be modeled as

$$P_{T\,d}(\boldsymbol{\upsilon}) = P_{Td}^{\mathrm{fix}} + B\sum_{k=1}^{K} \xi_k^{\mathrm{MS}} S_{edk}(\boldsymbol{\upsilon}) + \sum_{m=1}^{M} \left(\frac{\tau_d}{\tau_c} \frac{P_m^{\mathrm{tx}}(\boldsymbol{\upsilon})}{\alpha_m^{\mathrm{AP}}} + B\xi_m^{\mathrm{FH}} S_{ed}(\boldsymbol{\upsilon}) \right), \tag{50}$$

for the DL payload data transmission phase, and as

$$P_{T\,u}(\boldsymbol{\omega}) = P_{Tu}^{\mathrm{fix}} + \sum_{k=1}^{K} \frac{\tau_u}{\tau_c} \frac{P_u \omega_k}{\alpha_k^{\mathrm{MS}}} + B\sum_{m=1}^{M} \left(\xi_m^{\mathrm{AP}} + \xi_m^{\mathrm{FH}} \right) S_{eu}(\boldsymbol{\omega}), \tag{51}$$

for the UL payload data transmission phase, with

$$P_{Tl}^{\mathrm{fix}} = \frac{\tau_l}{\tau_c} \left[\sum_{k=1}^{K} P_{k\,l}^{\mathrm{MS,fix}} + \sum_{m=1}^{M} \left(P_m^{\mathrm{FH,fix}} + P_{m\,l}^{\mathrm{AP,fix}} + L_A P_{m\,l}^{\mathrm{AP,chain}} \right) + \sum_{m=1}^{M} \left(P_{m\,\mathrm{sleep}}^{\mathrm{FH,fix}} + P_{m\,\mathrm{sleep}}^{\mathrm{AP,fix}} + L_A P_{m\,\mathrm{sleep}}^{\mathrm{AP,chain}} \right) \right] \tag{52}$$

where l has been used as a token to represent either the DL ($l = d$) or the UL ($l = u$). As stated by Desset *et al.* in [50], although this simple linear model is not designed to provide very accurate absolute figures, it will enable a fair comparison among different on/off switching strategies for green cell-free massive-MIMO networking.

C. ENERGY EFFICIENCY

The energy efficiency during the DL and UL payload data transmission phases can be expressed as

$$E_{ed}(\boldsymbol{\upsilon}) = \frac{B S_{ed}(\boldsymbol{\upsilon})}{P_{T\,d}(\boldsymbol{\upsilon})} \tag{53}$$

and

$$E_{eu}(\boldsymbol{\omega}) = \frac{B S_{eu}(\boldsymbol{\omega})}{P_{T\,u}(\boldsymbol{\omega})}, \tag{54}$$

respectively. We can also define a *weighted* energy efficiency metric as

$$E_e(\boldsymbol{\upsilon}, \boldsymbol{\omega}) = (1-\mu) E_{ed}(\boldsymbol{\upsilon}) + \mu E_{eu}(\boldsymbol{\omega}), \tag{55}$$

where $0 \leq \mu \leq 1$ is a weighting coefficient allowing for the control of a trade-off between DL and UL energy efficiencies.

IV. AP SWITCHING STRATEGIES BASED ON GOODNESS-OF-FIT

Optimal ASO strategies aim at activating the subset of M_A APs providing the maximum energy efficiency. Determining the optimal subset of APs, however, is an NP-hard problem that calls for the evaluation of the performance provided by all possible combinations of M_A out of M APs. Hence, the implementation of computationally feasible selection strategies will only be possible by relying on the development of heuristic suboptimal algorithms. In the following we describe some heuristic ASO strategies that are based on the GoF theory. Under ideal conditions, the set of selected APs should be adapted to scenario variations due to, among others, changes in the number and/or location of MSs or changes in the geographical distribution of shadow fading. In most practical situations, however, these variations occur too quickly so as to allow the implementation of realistic ASO schemes that can adapt to them. The GoF-based ASO strategies have been specifically designed to cope with long-term non-uniform spatial traffic densities. In particular, it seems intuitively satisfactory to try to match the spatial distribution of active APs to that of the MSs in the network in an attempt to selectively

activate that parts of the network where most likely active users are located. This is the rational behind the GoF methods presented next. The performance provided by these ASO schemes will be benchmarked against that provided by three of the ASO strategies previously proposed in [25], that have been suitably adapted to the mmWave scenario:

- *random selection ASO (RS-ASO)*: Active APs are randomly selected and the only parameter that is optimized in trying to maximize the energy efficiency of the network is the number of active APs as a function of the number (or spatial density) of MSs in the serviced area. The energy efficiency performance improvement provided by the elementary RS-ASO strategy will serve as a lower bound on the performance improvement any other sensible ASO scheme may bring along.
- *minimum propagation losses-aware ASO (MPL-ASO)*: The set $\mathcal{M}^A$ of active APs is selected based on large-scale propagation losses between APs and MSs. For those cases in which $M_A \geq K$, the algorithm selects (in an ordered manner) the group of M_A APs showing the minimum propagation losses to the K MSs in the network. For those cases in which $M_A < K$, instead, a set of M_A *virtual* MSs is first generated by relying on the k-means clustering method and the previously described procedure is then applied to the virtual MSs to select the set of active APs. A detailed explanation of this ASO strategy can be found in [25]. It is important to note at this point that the pace at which the APs would have to be switched on/off under this strategy would be so high that it would be hardly implementable in practice.
- *optimal energy efficiency-based greedy ASO (OG-ASO)*: This is an iterative greedy algorithm that, starting with the M available APs in the first iteration, in the ith iteration of the algorithm evaluates the $(M+1-i)$ possible configurations of $(M-i)$ active APs resulting from switching off one of them, and selects the configuration maximizing the energy efficiency. The algorithm iterates until obtaining the configuration of active APs maximizing the energy efficiency of the network. The energy efficiency performance improvement provided by this *unrealistic* ASO strategy will serve as an upper bound against which to compare the performance provided by the other ASO schemes.

A. MOTIVATION FOR GOODNESS-OF-FIT

When a particular probability distribution has been specified to model a random phenomenon (such as the spatial distribution of MSs) the validity of the specified or assumed distribution model may be statistically verified or disproved by using GoF tests [54]–[58]. In this context, we can use GoF techniques to determine which APs should be turned on or off in such a way that the resulting AP distribution matches the non-uniform MS distribution.

As previously described in Section II-A, we consider that the target region is tessellated in a regular grid of N_X by N_Y rectangular pixels, and the probability density of MSs on pixel (x, y) is denoted as $f_{x,y}^{MS}$. For a given set $\mathcal{M}^A$ of active APs, we can determine an *estimate* of the probability density of APs $f_{x,y}^{AP}$ on this particular pixel as

$$f_{x,y}^{\mathrm{AP}} = \frac{M_A^{(x,y)}}{M_A}, \tag{56}$$

where $M_A^{(x,y)}$ is the number of active APs on pixel (x, y), and M_A is the number of active APs in the target region. Thus, relying on (56), the GoF techniques can establish a link between the spatial distribution of MSs (i.e., $f_{x,y}^{\mathrm{MS}}$ values) and the spatial distribution of active APs (i.e., $f_{x,y}^{\mathrm{AP}}$ values). In particular, in the following subsections three novel ASO strategies are proposed: two of them are based on widely applied GoF techniques [54], namely, the Chi-square test, and the Kolmogorov-Smirnov test, and the third one is based on the concept of *statistical energy*, described by Aslan and Zech in [59]. The optimal number of active APs (under any of the proposed GoF-based ASO strategies) when serving a given amount of MSs would be the one providing the maximum energy efficiency.

B. CHI-SQUARE BASED ASO

The Chi-square test (ChiS) is closely connected to a least square fit between the observed normalized frequencies of APs per pixel $\{f_{x,y}^{\mathrm{AP}}\}_{\forall(x,y)}$ with the corresponding theoretical probability densities $\{f_{x,y}^{\mathrm{MS}}\}_{\forall(x,y)}$. Given a set $\mathcal{M}^A$ of selected active APs, the GoF metric implemented by the ChiS test can be expressed as

$$D_{\mathcal{M}^A}^{(\mathrm{ChiS})} = \sum_{x=1}^{N_X}\sum_{y=1}^{N_Y} \frac{\left(f_{x,y}^{\mathrm{AP}} - f_{x,y}^{\mathrm{MS}}\right)^2}{f_{x,y}^{\mathrm{MS}}}. \tag{57}$$

The lower the value of $D_{\mathcal{M}^A}^{(\mathrm{ChiS})}$ the better the GoF between the spatial distributions of MSs and active APs. Hence, by using (57) the optimal ChiS-based ASO (ChiS-ASO) algorithm would be the one selecting the APs whose corresponding $D_{\mathcal{M}^A}^{(\mathrm{ChiS})}$ value is minimum. Having a large number of APs in the network, however, NP-hardness forbids the implementation of a brute force algorithm to solve this optimization problem. Consequently, an iterative ChiS-ASO algorithm is proposed that, starting with a set containing all the APs in the network, in each iteration switches-off the single AP leading to the minimum $D_{\mathcal{M}^A}^{(\mathrm{ChiS})}$ value when removed.

C. KOLMOGOROV-SMIRNOV BASED ASO

The ChiS test just described is an important special case of the power-divergence statistic [58] and is computationally simple. Unfortunately, however, it also has a serious drawback: as it neglects the correlation between adjacent elements of the histogram (i.e., between adjacent pixels), it exhibits a rather poor performance in detecting slowly varying deviations between both the analytical and predicted statistical distributions [55]. Furthermore, the ChiS scheme requires of a

large number of intervals and/or samples, that is, it requires of large M_A, N_X and N_Y values. In this regard, the Kolmogorov-Smirnov test (KS)-based approach has some advantages over the ChiS test. In particular, with the KS-based strategy the problem associated with small number of intervals and/or samples would not be an issue [55]. Moreover, the KS-based ASO (KS-ASO) algorithm takes into consideration the possible correlations between adjacent elements (or pixels) by using the cumulative distribution functions (CDFs) in calculating the discrepancy metric.

For the one-dimensional case, the KS test considers simply the largest absolute difference between the two CDFs as a measure of misfit. In this case, the result of the test is independent of the direction of ordering of the data, that is, it is independent of whether we consider the cumulative probabilities $P(x > X)$ or $P(x < X)$. In a multi-dimensional case, however, defining the CDF as $P(x < X, y < Y, \ldots)$ is ambiguous since the directions in which we choose to order the different random variables are arbitrary. In fact, in an n-dimensional case there are $2^n - 1$ independent ways of defining the CDF. A straightforward way to avoid the dependency of the KS test on the particular orderings chosen is to specify the discrepancy metric as the largest absolute difference between empirical and theoretical CDFs when all possible ordering combinations (i.e., 2^n) are considered [60], [61]. In our particular two-dimensional case, this corresponds to recognizing that the statistical descriptions of the spatial location of both APs and MSs in all four quadrants of the plane defined by $(x < X, y < Y)$, $(x < X, y > Y)$, $(x > X, y < Y)$ and $(x > X, y > Y)$ are equally valid, and that the discrepancy metric can be obtained as the largest of the four differences in empirical and theoretical CDFs. This can be mathematically expressed as

$$D_{\mathcal{M}^A}^{(\mathrm{KS})} = \max_{i\in\{1,2,3,4\}} \left\{ \max_{x,y} \left| F_{x,y}^{\mathrm{AP},i} - F_{x,y}^{\mathrm{MS},i} \right| \right\}, \qquad (58)$$

where the theoretical and empirical CDFs describing the spatial distributions of MSs and APs, respectively, in the four quadrants of the plane can be obtained from $f_{x,y}^{\mathrm{MS}}$ in (3) and $f_{x,y}^{\mathrm{AP}}$ in (56) as

$$F_{x,y}^{CE,1} = \sum_{i=1}^{x}\sum_{j=1}^{y} f_{i,j}^{CE}, \qquad F_{x,y}^{CE,2} = \sum_{i=1}^{x}\sum_{j=y}^{N_Y} f_{i,j}^{CE},$$

$$F_{x,y}^{CE,3} = \sum_{i=x}^{N_X}\sum_{j=1}^{y} f_{i,j}^{CE}, \qquad F_{x,y}^{CE,4} = \sum_{i=x}^{N_X}\sum_{j=y}^{N_Y} f_{i,j}^{CE}, \qquad (59)$$

with CE denoting a token used to represent one of the communication ends, either the AP (i.e., $CE = \mathrm{AP}$) or the MS (i.e., $CE = \mathrm{MS}$).

Using, once again, a greedy strategy starting with a set containing all the APs in the network, an iterative KS-ASO algorithm switches-off, in each iteration, the single AP resulting in the minimum $D_{\mathcal{M}^A}^{(\mathrm{KS})}$ value when removed.

D. STATISTICAL ENERGY-BASED ASO

In [59], Aslan and Zech introduce the concept of *statistical energy* as a tool for multivariate GoF tests. Similar to what is done when dealing with electric charge distributions, where charges of opposite sign are in a state of minimum energy when they are equally distributed, they define the *statistical energy* of statistical distributions and use it to mathematically describe a GoF test.

For a given set $\mathcal{M}^A = \left\{m_1^A, \ldots, m_{M_A}^A\right\}$ of active APs, the statistical energy-based test aims at comparing the sample of spatial locations of these APs (which follow an unknown pdf) to the reference theoretical spatial pdf $f_{x,y}^{\mathrm{MS}}$ of the MSs. To this end, let us first define the location (on a complex plane) of AP m_l^A, for all $l \in \{1, \ldots, M_A\}$, as $p_{m_l^A}^{\mathrm{AP}}$, with *charge* $1/M_A$. Furthermore, let us use the locations $p_{x,y}$ of pixels (x, y) for all $x \in \{1, \ldots, N_X\}$ and $y \in \{1, \ldots, N_Y\}$, which conform to the spatial pdf $f_{x,y}^{\mathrm{MS}}$ in (3). Using these components, the statistical energy test statistic in [59, (3.1)] can be adapted to our problem at hand as

$$D_{\mathcal{M}^A}^{(LSE)} = \frac{1}{M_A(M_A-1)} \sum_{l=1}^{M_A}\sum_{n=l+1}^{M_A} R_{\log}\left(\left|p_{m_l^A}^{AP} - p_{m_n^A}^{AP}\right|\right) - \frac{1}{M_A}\sum_{x=1}^{N_X}\sum_{y=1}^{N_Y} f_{x,y}^{MS} \sum_{l=1}^{M_A} R_{\log}\left(\left|p_{x,y} - p_{m_l^A}^{AP}\right|\right), \qquad (60)$$

where the logarithmic distance $R_{\log}(r) = -\ln(r + \varepsilon)$, with $\varepsilon = 1/(2\, N_X N_Y \max\{f_{x,y}^{MS}\})$, is used because, as shown in [59, Fig.2], it provides a reasonably good performance in scenarios with very dissimilar spatial distributions (background contaminations in the notation used by Aslan and Zech in [59]). Again, a greedy strategy is implemented that, starting with a set containing all the APs in the network, implements an iterative logarithmic statistical energy ASO (LSE-ASO) algorithm that switches-off, in each iteration, the single AP resulting in the minimum $D_{\mathcal{M}^A}^{(LSE)}$ value when removed.

V. NUMERICAL RESULTS

This section presents a comprehensive set of numerical results to qualitatively and quantitatively assess the performance of the proposed ASO strategies in a cell-free mmWave massive MIMO context in terms of both energy and spectral efficiencies, and also overall power consumption. Particular attention is paid to the effects caused by modifying the RF infrastructure at the APs and the consequences of changing the density of MSs per area unit and their corresponding spatial distribution. As generally done in most cell-free background literature, an scenario is considered where APs are initially uniformly distributed at random within a square coverage area of side D whose boundaries are wrapped around, thus effectively simulating the operation of a network without boundary effects. Unlike most previous works, however, the positions of the MSs follow a non-uniform distribution as previously explained in Section II-A.

The main parameters used throughout all simulations in this section are collected in Table 2. These parameters have

TABLE 2. Summary of default simulation parameters.

Parameters	Value
Carrier frequency: f_0	28 GHz
Bandwidth: B	20 MHz
Side of the square coverage area: D	500 m
AP/MS antenna height: $h_{\text{AP}}/h_{\text{MS}}$	10/1.65 m
Noise figure at the MS: NF_{MS}	9 dB
Noise figure of the LNA at the AP: NF_{LNA}	1.6 dB
Gain of the LNA at the AP: G_{LNA}	22 dB
Phase splitters attenuation of the at the AP: L_{PS}	3 dB
Power combiner attenuation at the AP: $L_{\text{PC}_{in}}$	3 dB
Noise figure of the RF chain at the AP: NF_{RF}	7 dB
AP maximum transmit power: P_d	200 mW
MS maximum transmit power: $P_u = P_p$	100 mW
Antenna configuration at each AP: $N_x \times N_y$	8×1
Minimum separation between antenna elements	$\lambda/2$
Coherence interval length: τ_c	200 samples
Training phase length: τ_p	20 samples
Parameters for MS condition: a_{out}, b_{out}, a_{LOS}	1/30, 5.2, 1/67.1
Pathloss parameters Case LOS: α, β, σ_χ	61.34, 2.10, 4.0 dB
Pathloss parameters Case NLOS: α, β, σ_χ	61.34, 3.19, 8.2 dB
Shadow fading decorrelation distance: d_{dcorr}	9 m
Shadow fading correlation among APs:	0.5
Ricean K-factor distribution: μ_K/σ_K	9/5 dB
Number of clusters LOS/NLOS: C_{mk}	12/19
Number of paths per cluster LOS/NLOS: P_{mk}	20/20
Azimuth angular spread (AP) LOS/NLOS: $\lambda_{\text{rms}}^{-1}$	$3^o/10^o$
Elevation angular spread: $\lambda_{\text{rms}}^{-1}$	7^o
Cluster power fraction parameters: r_τ/ζ	3/4
Power amplifier efficiency: $\alpha_m^{\text{AP}}/\alpha_k^{\text{MS}}$	0.39/0.3
Power per AP/MS/FH traffic: $\xi_m^{\text{AP}}/\xi_k^{\text{MS}}/\xi^{\text{FH}}$	0.25/0.25/0.25 $\frac{\text{W}}{\text{Gbps}}$
AP fixed power: $P_{m\,d}^{\text{AP,fix}}/P_{m\,u}^{\text{AP,fix}}$	8/6 W
AP fixed power per RF chain: $P_{m\,d}^{\text{AP,chain}}/P_{m\,u}^{\text{AP,chain}}$	0.2/0.15 W
AP fixed power (sleep): $P_{m\,\text{sleep}}^{\text{AP,fix}}$	0.8 W
AP fixed power per RF chain (sleep): $P_{m\,\text{sleep}}^{\text{AP,chain}}$	0.02 W
MS fixed power: $P_{k\,d}^{\text{MS,fix}}/P_{k\,u}^{\text{MS,fix}}$	1/0.75 W
FH fixed power: $P_m^{\text{FH,fix}}$	5 W
FH fixed power (sleep): $P_{m\,\text{sleep}}^{\text{FH,fix}}$	0.5 W

been borrowed from a variety of prior research works (see, for instance, [4], [16], [29], [31], [48], [51], [53]). Results shown next have been obtained using the heuristic power allocation introduced by Nayebi *et al.* in [5, eq. (21)] (i.e., $\upsilon_k = P_d / \left(\max_m \sum_{k'=1}^{K} \theta_{mk'}\right)$ for all k) for the DL case, which can be deemed as a computationally simple approximation to the max-min power allocation approach, and a full-power transmission strategy (i.e., $\omega_k = 1$ for all k) for the UL case. Nonetheless, it should be stressed that the proposed framework is indeed applicable to any other power allocation policy (such as those presented in, for example, [2], [3], [5]). Finally, and following the work in [16], a balanced random pilot assignment scheme is implemented whereby MSs are allocated pilot sequences that are sequentially and cyclically selected from the ordered set of available orthogonal pilots.

A. IMPACT OF THE ASO STRATEGY

We start by assessing in this subsection the performance achieved by each of the proposed and considered ASO strategies. Towards this end, Fig. 3 shows the impact of the ASO strategy by depicting the overall average weighted energy efficiency as a function of the number of active APs when a DL/UL weighting coefficient $\mu = 0.5$ (i.e., DL and UL are given the same importance) has been enforced. It has been assumed that the total number of APs in the system is $M = 100$ and each of them is equipped with an 8×1 uniform linear array (ULA) of vertical half-wave dipoles and $L = 4$ RF chains. The results in this figure have been obtained when simultaneously serving $K = 16$ MSs. As anticipated, the energy efficiency achieved by the RS-ASO and OG-ASO schemes act, respectively, as lower- and upper-bounds on the performance attained by any of the other considered ASO strategies. Note that the proposed ASO schemes can be classified in three groups as a function of the system state information they manage. In particular, the RS-ASO scheme would be the only member in the first group, comprising those ASO strategies that are completely unaware of the network state and thus make blind AP switch-on/off decisions. The second group, comprising the goodness-of-fit techniques-based ASO schemes (i.e., the ChiS-ASO, the KS-ASO and the LSE-ASO), assume that the spatial distribution of MSs on the service coverage area is known, and make only use of *very large-scale* system-state information in the form of the geographical location of the APs. It is remarkable the rather significant performance improvement provided by these methods over the pure RS-ASO given their reliance of such a *coarse* network-state information. Note how the achievable energy efficiency increases the more APs are switched-off up to a certain point where an optimum is reached which, for this particular number of MSs, is located around $M_A = 24$, 20 and 18 active APs for ChiS-ASO, KS-ASO and LSE-ASO strategies, respectively, which shows a significant deviation from the optimum reached when relying on RS-ASO, located around $M_A = 34$ APs.

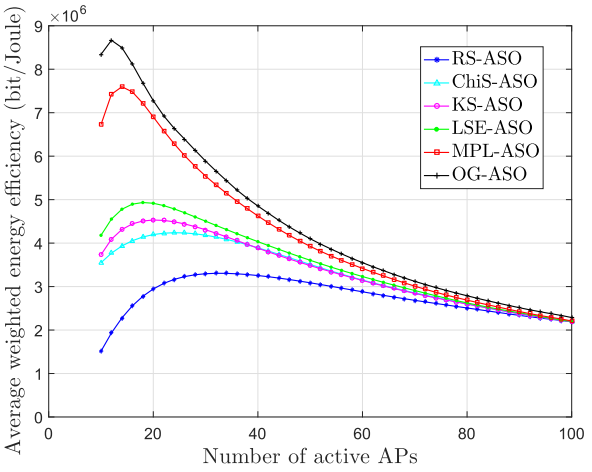


FIGURE 3. Impact of the ASO strategy on the average (equally) weighted energy efficiency as a function of the number of active APs (weighted UL/DL scenario with $\mu = 0.5$).

The MPL-ASO and OG-ASO strategies make up the third group, characterized by a certain degree of knowledge of the *short-term* network-state information in the form of the large-scale propagation losses between APs and MSs. Note that

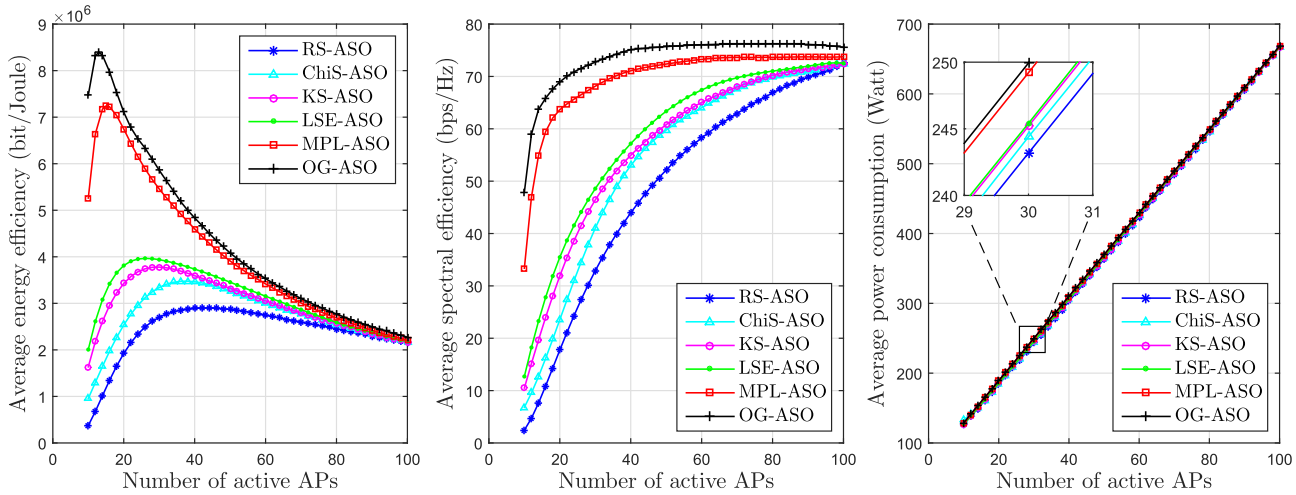


FIGURE 4. Impact of the ASO strategy on the DL average energy efficiency, spectral efficiency and power consumption as a function of the number of active APs.

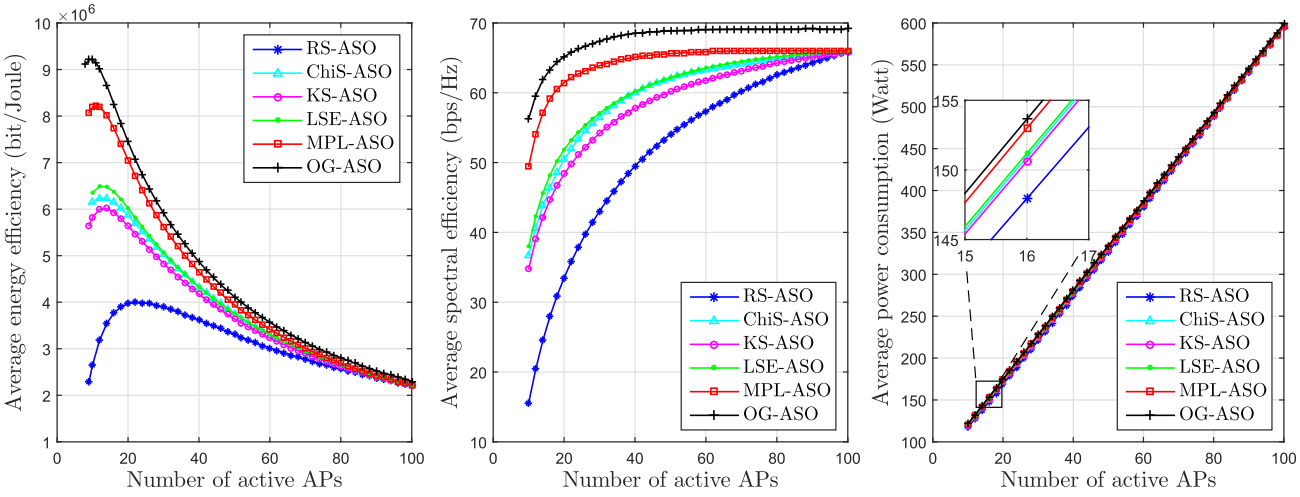
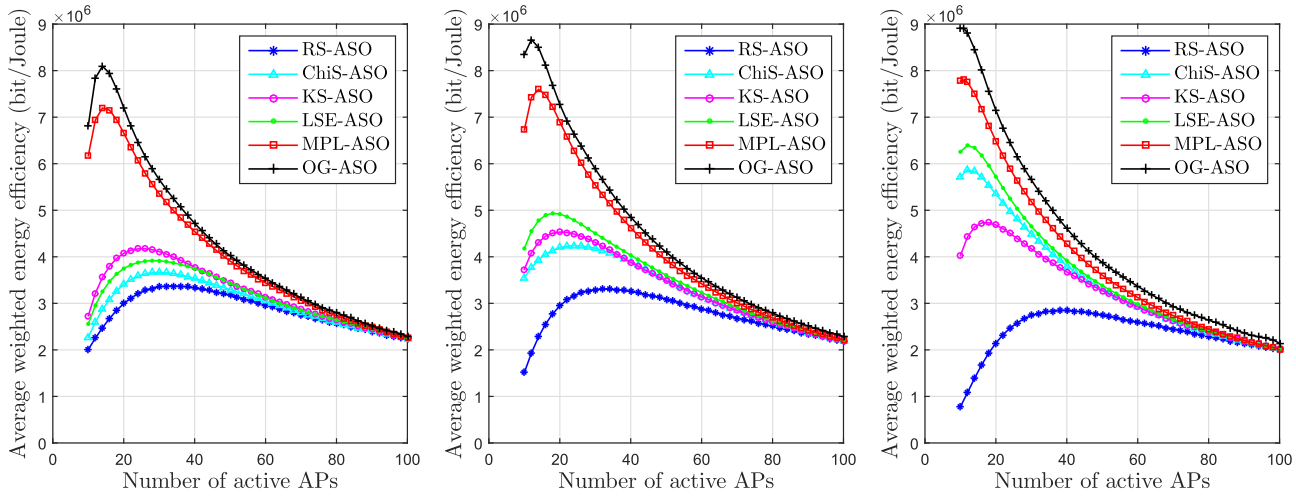


FIGURE 5. Impact of the ASO strategy on the UL average energy efficiency, spectral efficiency and power consumption as a function of the number of active APs.

these losses are tightly connected to the instantaneous positions of the users over the coverage area. The MPL-ASO strategy dynamically adapts to *short-term* variations of the spatial distribution of MS and, as shown in Fig. 3, greatly outperforms the ASO strategies in the first and second groups. The energy efficiency provided by this strategy increases when switching-off some of the APs until only M_A =14 APs are left active. Finally, the OG-ASO scheme assumes the complete knowledge of all long-term network-state information necessary to calculate the achievable energy-efficiency, including, among others, the channel spatial correlation matrices, the power control matrices or the power consumption metrics and it is seen to outperform the rest of techniques. However, and rather strikingly, the energy-efficiency performance gap between this *idealistic* approach and the much simpler MPL-ASO scheme is modest and, moreover, the optimum number of APs to be left active virtually coincides (M_A =14 for MPL-ASO *vs* M_A =13 for OG-ASO).

The energy efficiency, spectral efficiency and power consumption *versus* the number of active APs is presented in Figs. 4 and 5 for each of the considered schemes and for both DL (i.e., μ =0) and UL (i.e., μ =1), respectively. Note how now, when considering a pure DL setting (Fig. 4), the optimal number of active APs necessary to serve K = 16 MSs is M_A =44, 38, 30 and 26 APs for the RS-ASO, ChiS-ASO, KS-ASO and LSE-ASO strategies, respectively, and M_A =15 and 13 APs for MPL-ASO and OG-ASO strategies, respectively. In contrast, when focusing on the UL case (Fig. 5), the optimal number of active APs necessary to serve K = 16 MSs is M_A =23, 14, 13 and 12 APs for the RS-ASO, ChiS-ASO, KS-ASO and LSE-ASO strategies, respectively, and M_A =10 and 8 APs for MPL-ASO and OG-ASO strategies, respectively. Jointly considering DL and UL it is easy to conclude that, irrespective of the ASO in use, the DL requires of significantly more infrastructure to be active when compared to the UL, roughly the double



(a) μ =0.5, $\sigma^{\mathcal{S}}$= 0.5297 (b) μ =0.5, $\sigma^{\mathcal{S}}$= 2.1188 (c) μ =0.5, $\sigma^{\mathcal{S}}$= 8.4752

FIGURE 6. Average energy efficiency as a function of the number of active APs under scattered (homogeneous), urban and concentrated (hot-spot) spatial distributions of MSs (weighted UL/DL scenario with $\mu = 0.5$).

number of active APs if energy efficiency optimality is to be preserved. In fact, it can be generally concluded that for any fixed number of APs, the energy efficiency of the UL always exceeds that of the DL. This is basically due to two reasons. Firstly, the average power consumption metrics in the UL are considerably lower than the corresponding DL ones. Secondly, for an optimal number of active APs, the use of full power transmission in the UL provides a clear advantage, in terms of spectral efficiency, with respect to the constrained power control transmission implemented in the DL. If a max-min power control strategy was implemented in both UL and UL, an almost identical spectral efficiency performance would be obtained in both cases (see, for instance, results presented in [16]), and the energy efficiency advantage shown by the UL segment would only be due, in this case, to the lower fixed power consumption.

B. IMPACT OF THE DISTRIBUTION OF MSs

In order to seize how the concentration of MSs influences the decision to choose the most appropriate scheme to maximize the energy efficiency, the effect that parameter $\sigma^{\mathcal{S}}$ has on the network performance is now assessed. In particular, results shown in Fig. 6 are presented for $\sigma^{\mathcal{S}} = 2.1188$, which corresponds to an *urban common* distribution of MSs and that can serve as a baseline against which to compare more *concentrated* distributions ($\sigma^{\mathcal{S}} = 8.4752$) indicating the presence of *hot spots*, and more *scattered* ones ($\sigma^{S} = 0.5297$), representative of more *homogeneous* suburban environments. Throughout this figure a symmetric UL/DL split is considered ($\mu = 0.5$).

The first and foremost effect worthwhile pointing out is that ASO strategies based on GoF offer an energy efficiency performance that greatly exceeds the one attained under the random approach. In fact, note how, as expected, although the performance of the pure RS-ASO algorithm is quite acceptable under very homogeneous spatial traffic distributions, it exhibits a clear degradation as the MS distribution becomes more heterogeneous (i.e., with increasing $\sigma^{\mathcal{S}}$). Among the GoF-based schemes, the *homogeneous* spatial distribution is best discriminated by the KS-ASO strategy and the ChiS-ASO scheme just offers a slight improvement with respect to the pure random approach. The LSE-ASO algorithm is also quite powerful under this circumstances and clearly outperforms the ChiS-ASO scheme. As the concentration of MSs increases, the behaviours of the ChiS-ASO and KS-ASO strategies invert. In fact, the sensitivity of the KS-ASO scheme to changes in the spatial distribution of MSs proves to be very poor. Remarkably, however, although the discrimination power of the ChiS-ASO algorithm improves when dealing with *concentrated* spatial distributions of MSs, this ASO strategy is clearly outperformed by the LSE-ASO approach. Summarizing, even though the KS-ASO scheme would be the GoF-based strategy of choice in scenarios showing a high degree of homogeneity in the spatial distribution of MSs, the discrimination power of the LSE-ASO approach is the one showing less dependence on the traffic spatial distribution, thus making it the most versatile of them. These results are very consistent with those presented by Aslan and Zech in [59].

The MPL-ASO and OG-ASO techniques, owing to their reliance on far more detailed network information, show a considerable improvement in energy-efficiency performance over the rest of approaches. Moreover, similar to the GoF-based techniques, and given their inherent capability to more adequately respond to the spatial distribution of MSs, they also reflect a very significant improvement with increasing $\sigma^{\mathcal{S}}$.

Building on the remarks just made, and based on the degree of use of network-state information, performance and complexity, it can be concluded that the KS-ASO and LSE-ASO,

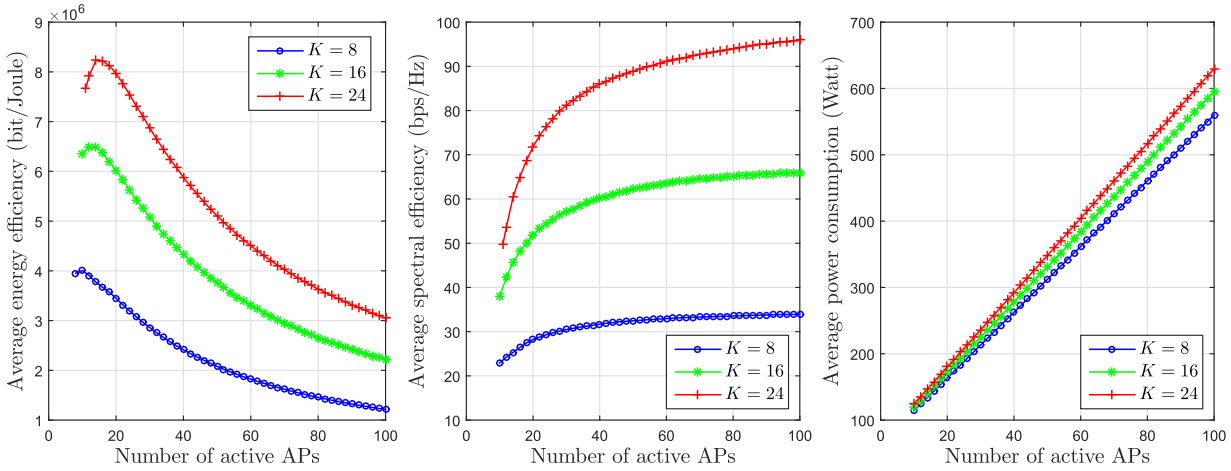


FIGURE 7. Impact of the number of MSs on the UL average energy efficiency, spectral efficiency and power consumption as a function of the number of active APs under the LSE-ASO strategy.

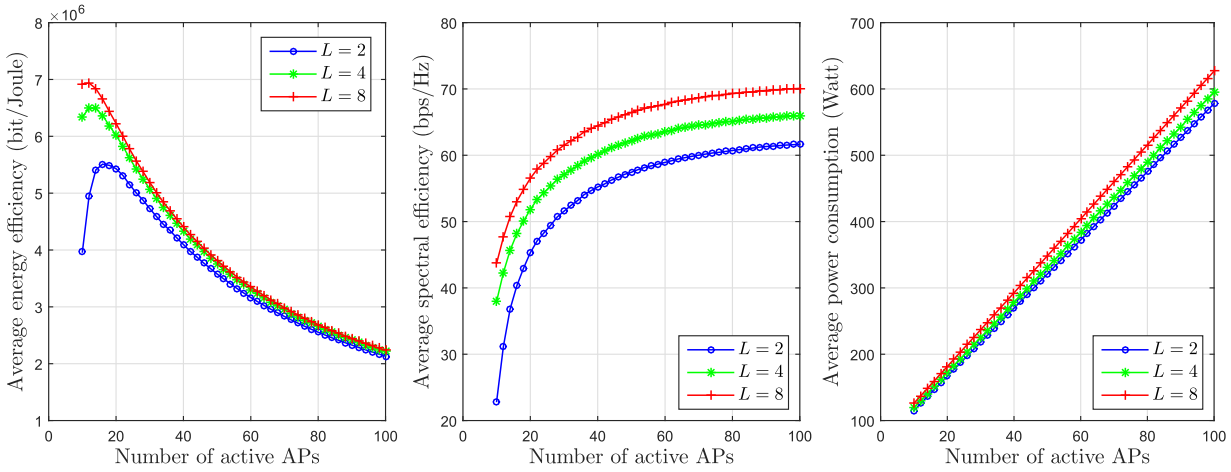


FIGURE 8. Impact of the RF infrastructure used at the APs on the UL average energy efficiency, spectral efficiency and power consumption as a function of the number of active APs under the LSE-ASO strategy.

among all the proposed ASO strategies, represent the most adequate schemes to be implemented in a cell-free mmWave massive MIMO system. In particular, KS-ASO has shown to be very effective in exploiting mild deviations from MS distribution homogeneity ($\sigma^{\mathcal{S}}0.5297$), whereas LSE-ASO provides the best discrimination power when dealing with spatial distributions of MSs showing a marked heterogeneity. Owing to its very good performance across a variety of $\sigma^{\mathcal{S}}$ values, indicative of its robustness, results presented over the next subsections will focus on the use of the LSE-ASO scheme. Furthermore, without loss of essential generality, only the optimization of the UL segment (i.e., $\mu = 1$) will be considered (conclusions drawn by using other ASO strategies and any other weighting coefficient μ would be qualitatively equivalent).

C. IMPACT OF THE NUMBER OF MSs IN THE NETWORK

Figure 7 studies the impact the number of MSs has on the UL average energy efficiency, spectral efficiency and power consumption as a function of the number of APs. As it can be observed, increasing the number of MSs in the network results in an increase in both the average spectral efficiency and the average power consumption. However, the dissimilar increments induced in these two metrics result in quite unalike effects on the average energy efficiency of the network. In particular, fixing the number of active APs, when increasing the network load (i.e., more MSs), results in average spectral efficiency increments that more than compensate for the increase in average power consumption, hence raising the average energy efficiency. On the contrary, fixing the number of MSs, increments in the number of active APs translate into small improvements in spectral efficiency and a considerable raise of the consumed power, thus resulting in a substantial deterioration of the average energy efficiency of the network. Therefore, when aiming at a high energy efficiency of the a cell-free mmWave massive MIMO network it is of prime concern that the number of active APs is appropriately adapted to the number of MSs to be serviced. This insight is in fact reinforced by noting that the optimal

number of active APs increases with the number of MSs in the network.

D. IMPACT OF THE RF INFRASTRUCTURE USED AT THE APs

To understand how the RF infrastructure used at the APs influences the performance of the system, Fig. 8 represents the energy efficiency, spectral efficiency and power consumption *versus* the number of active APs assuming that each of them is equipped with an 8×1 linear uniform array and use an analog precoder with $L = 2$, 4 and 8 RF chains. Note that the latter case indeed corresponds to a system with fully digital processing capability. Results shown next have been obtained assuming the use of a LSE-ASO strategy, and the availability of $M = 100$ APs to serve $K = 16$ MSs. Naturally, both the average spectral efficiency and the power consumption increase with the number of available RF chains. Since the number of active RF chains at each of the APs in the network is equal to $L_A = \min\{K, L\}$, increasing the number of available RF chains is always beneficial for scenarios where $K \geq L$. Furthermore, note how, as Fig. 7 reveals, the optimum number of active APs to maximize energy efficiency decreases with the number of RF chains available at each AP. In particular, for the scenario under consideration, with $L = 2$ RF chains the optimal number of active APs is equal to $M_A^* = 16$, whereas using $L = 8$ RF chains at each of the APs, the corresponding optimal number of active APs is equal to $M_A^* = 12$.

VI. CONCLUSION

This paper has presented a comprehensive analytical framework for the evaluation of the energy efficiency of AP sleep-mode techniques for cell-free mmWave massive MIMO networks with non-uniform spatial traffic density. Based on this framework, different ASO strategies have been proposed whose goal is to dynamically turn on/off some of the APs in accordance with metrics related to the spatial distribution of MSs in the network with the objective of maximizing the energy efficiency. Towards this end, a realistic model to describe a non-uniform distribution of MSs has been included in the analysis that serves to capture the spatial traffic heterogeneity.

Aside from revisiting known ASO algorithms in the new context, three novel schemes based on goodness-of-fit techniques (i.e., ChiS-ASO, KS-ASO and LSE-ASO) have been introduced whose rationale is to try to match the spatial distribution of active APs to the one of the MSs. Additionally, one more technique, termed MPL-ASO, that relies on the AP-to-MS propagation losses has also been postulated. Remarkably, this new family of GoF-based ASO strategies (in particular KS-ASO and LSE-ASO) has been shown to perform considerably better than a pure random procedure while relying only on very large scale information in the form of estimates of the spatial distribution of MSs and spatial location of the APs. In particular, KS-ASO has shown to be very effective in exploiting mild deviations from MS distribution homogeneity, whereas LSE-ASO provides the best discrimination power when dealing with spatial distributions of MSs showing a marked heterogeneity and, furthermore, its performance is very robust against changes in the spatial distribution of traffic. In turn, the MPL-ASO algorithm, while requiring some large-scale information (i.e., AP-MS large-scale fadings), can neglect extra information the *optimum* OG-ASO technique requires (i.e., spatial correlation matrices, power control matrices, power consumption metrics) with only a small performance penalty while being considerably less complex. Results have shown that increasing the network load (more MSs) implies activating more APs to attain the optimum point of operation in terms of energy efficiency. The RF infrastructure at the APs is also seen to play a key role. In particular, the average energy efficiency increases as the number of active RF chains used at the APs grows but, interestingly, this in turns allows optimum operation with a smaller number of active APs.

Future work will concentrate on the use of more sophisticated power control strategies and hybrid analog-digital precoding stages, the study of more reactive ASO algorithms as well as its implementation issues, and also, the impact the use of finite-capacity fronthaul links between the APs and the CPU may have on the current analytical framework.

REFERENCES

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE SPAWC*, Jun. 2015, pp. 201–205.

[2] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[4] E. Bjornson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[5] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[6] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[7] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[8] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[9] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless Networks—With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[10] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 211–217, Apr. 2018.

[11] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart., 2018.

[12] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[13] X. Gao, L. Dai, S. Han, C.-L. I, 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, Apr. 2016.

[14] M. Alonzo and S. Buzzi, "Cell-free and user-centric massive MIMO at millimeter wave frequencies," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[15] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 651–663, Sep. 2019.

[16] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596–44612, Apr. 2019.

[17] A. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, Apr. 2015.

[18] P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: A survey," *IEEE Access*, vol. 5, pp. 11727–11758, Jun. 2017.

[19] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 803–826, 2nd Quart., 2015.

[20] F. Han, S. Zhao, L. Zhang, and J. Wu, "Survey of strategies for switching off base stations in heterogeneous networks for greener 5G systems," *IEEE Access*, vol. 4, pp. 4959–4973, Aug. 2016.

[21] H. Tabassum, U. Siddique, E. Hossain, and M. J. Hossain, "Downlink performance of cellular systems with base station sleeping, user association, and scheduling," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5752–5767, Oct. 2014.

[22] C. Jia and T. J. Lim, "Resource partitioning and user association with sleep-mode base stations in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3780–3793, Jul. 2015.

[23] X. Xu, C. Yuan, W. Chen, X. Tao, and Y. Sun, "Adaptive cell zooming and sleeping for green heterogeneous ultradense networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1612–1621, Feb. 2018.

[24] H. Jiang, S. Yi, L. Wu, H. Leung, Y. Wang, X. Zhou, Y. Chen, and L. Yang, "Data-driven cell zooming for large-scale mobile networks," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 156–168, Mar. 2018.

[25] G. Femenias, N. Lassoued, and F. Riera-Palou, "Access point switch ON/OFF strategies for green cell-free massive MIMO networking," *IEEE Access*, vol. 8, pp. 21788–21803, 2020.

[26] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," 2020, *arXiv:2002.01504*. [Online]. Available: http://arxiv.org/abs/2002.01504

[27] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.

[28] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[29] *Study on Channel Model for Frequency Spectrum Above 6 GHz (Release 14), Version 14.3.1*, document 3GPP TR 38.900, Jul. 2017.

[30] M. K. Samimi and T. S. Rappaport, "Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 3483–3489.

[31] C. T. Neil, M. Shafi, P. J. Smith, P. A. Dmochowski, and J. Zhang, "Impact of microwave and mmWave channel models on 5G systems performance," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6505–6520, Dec. 2017.

[32] O. Ozdogan, E. Bjornson, and E. G. Larsson, "Massive MIMO with spatially correlated Rician fading channels," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3234–3250, May 2019.

[33] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[34] S. Haghighatshoar and G. Caire, "Massive MIMO pilot decontamination and channel interpolation via wideband sparse channel estimation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8316–8332, Dec. 2017.

[35] D. Neumann, M. Joham, and W. Utschick, "Covariance matrix estimation in massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 863–867, Jun. 2018.

[36] K. Upadhya and S. A. Vorobyov, "Covariance matrix estimation for massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 546–550, Apr. 2018.

[37] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, Jun. 2014.

[38] R. Mendez-Rial, N. González-Prelcic, and R. W. Heath, Jr., "Adaptive hybrid precoding and combining in mmWave multiuser MIMO systems based on compressed covariance estimation," in *Proc. IEEE 6th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2015, pp. 213–216.

[39] S. Park, J. Park, A. Yazdan, and R. W. Heath, "Exploiting spatial channel covariance for hybrid precoding in massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3818–3832, Jul. 2017.

[40] R. Mai, T. Le-Ngoc, and D. H. N. Nguyen, "Two-timescale hybrid RF-baseband precoding with MMSE-VP for multi-user massive MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4462–4476, Jul. 2018.

[41] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.

[42] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, 2nd Quart., 2016.

[43] H. Q. Ngo, H. Tataria, M. Matthaiou, S. Jin, and E. G. Larsson, "On the performance of cell-free massive MIMO in Ricean fading," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 980–984.

[44] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[45] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[46] H. Yang and T. L. Marzetta, "Capacity performance of multicell large-scale antenna systems," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2013, pp. 668–675.

[47] G. Interdonato, H. Q. Ngo, E. G. Larsson, and P. Frenger, "On the performance of cell-free massive MIMO with short-term power constraints," in *Proc. IEEE 21st Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Oct. 2016, pp. 225–230.

[48] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[49] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2011, pp. 1–5.

[50] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez, H. Klessig, I. Gódor, M. Olsson, M. A. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of LTE base stations," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 2858–2862.

[51] E. Bjornson, L. Sanguinetti, and M. Kountouris, "Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 832–847, Apr. 2016.

[52] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[53] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871–1874, Aug. 2017.

[54] R. B. D'Agostino, *Goodness-of-Fit-Techniques*, vol. 68. Boca Raton, FL, USA: CRC Press, 1986.

[55] B. Aslan and G. Zech, "Comparison of different goodness-of-fit tests," 2002, *arXiv:math/0207300*. [Online]. Available: https://arxiv.org/abs/math/0207300

[56] D. L. Evans, J. H. Drew, and L. M. Leemis, "The distribution of the Kolmogorov–Smirnov, Cramer–Von Mises, and Anderson–darling test statistics for exponential populations with estimated parameters," *Commun. Statistics–Simulation Comput.*, vol. 37, no. 7, pp. 1396–1421, 2008.
[57] M. Williams, "How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics," *J. Instrum.*, vol. 5, no. 9, Sep. 2010, Art. no. P09004.
[58] T. R. Read and N. A. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Cham, Switzerland: Springer, 2012.
[59] B. Aslan and G. Zech, "Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding," *Nucl. Instrum. Methods Phys. Res. A, Accel. Spectrom. Detect. Assoc. Equip.*, vol. 537, no. 3, pp. 626–636, Feb. 2005.
[60] J. A. Peacock, "Two-dimensional goodness-of-fit testing in astronomy," *Monthly Notices Roy. Astronomical Soc.*, vol. 202, no. 3, pp. 615–627, Mar. 1983.
[61] G. Fasano and A. Franceschini, "A multidimensional version of the Kolmogorov–Smirnov test," *Monthly Notices Roy. Astronomical Soc.*, vol. 225, pp. 155–170, Mar. 1987.

JAN GARCÍA-MORALES (Member, IEEE) received the degree in telecommunications and electronic engineering from the Central University of Las Villas (UCLV), Cuba, in July 2008, and the Ph.D. degree in information and communication technologies from the University of the Balearic Islands, Spain, in March 2018. In 2017, he spent a period of three months with the School of Engineering, The University of Edinburgh, Scotland, U.K., under a Research-Stay, and the Cooperation Grant supported by the La Caixa Foundation. In 2019, he was a Postdoctoral Researcher with the UWICORE Laboratory, Miguel Hernández University (UMH), Elche, Spain, under the framework of the European H2020 AUTOWARE Research Project. He is currently a Postdoctoral Researcher with the Mobile Communications Group (MCG), University of the Balearic Islands. His Ph.D. degree was supported by MINECO, Spanish Government, FEDER, and CAIB, Govern Balear. He was on solutions for self-organizing cellular networks that support Industry 4.0 applications and scenarios with a RAN slicing focus and dedicated to the study of energy efficient solutions for cell-free massive MIMO networks. He received the Best Ph.D. students SANTANDER-UIB Award from the University of the Balearic Islands and the Santander Universidades Foundation.

GUILLEM FEMENIAS (Senior Member, IEEE) received the degree in telecommunication engineering and the Ph.D. degree in electrical engineering from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1987 and 1991, respectively. From 1987 to 1994, he was a Researcher with UPC, where he was an Associate Professor, in 1992. In 1995, he joined the Department of Mathematics and Informatics, University of the Balearic Islands (UIB), Mallorca, Spain, where he was a Full Professor, in 2010. He is currently Leads the Mobile Communications Group, UIB. He was a Project Manager with projects ARAMIS, DREAMS, DARWIN, MARIMBA, COSMOS, ELISA, and TERESA, funded by the Spanish and the Balearic Islands Governments. He was involved with several European projects, such as ATDMA, CODIT, and COST. He has published more than 100 journal and conference papers and some book chapters. His current research interests include digital communications theory and wireless communication systems, with particular emphasis on radio resource management strategies applied to 5G and 6G wireless networks. He has served on various the IEEE conferences, as a technical program committee member. He was also a Local Organizing Committee Member with the IEEE Statistical Signal Processing (SSP), in 2016. He served as the Publications Chair for the IEEE 69th Vehicular Technology Conference (VTC-Spring), in 2009. He was a recipient of the Best Paper Awards from the 2007 IFIP International Conference on Personal Wireless Communications and the 2009 IEEE Vehicular Technology Conference-Spring.

FELIP RIERA-PALOU (Senior Member, IEEE) received the B.S. and M.S. degrees in computer engineering from the University of the Balearic Islands (UIB), Mallorca, Spain, in 1997, the M.Sc. and Ph.D. degrees in communication engineering from the University of Bradford, U.K., in 1998 and 2002, respectively, and the M.Sc. degree in statistics from The University of Sheffield, U.K., in 2006. From May 2002 to March 2005, he was with the Philips Research Laboratories, Eindhoven, The Netherlands, as a Marie Curie Postdoctoral Fellow (European Union) and a Technical Staff Member. He worked on research programs related to wideband speech/audio compression and speech enhancement for mobile telephony. From April 2005 to December 2009, he was a Research Associate (Ramon y Cajal Program, Spanish Ministry of Science) with the Mobile Communications Group, Department of Mathematics and Informatics, UIB, where he has been an Associate Research Professor, I3 Program, Spanish Ministry of Education, since January 2010. His current research interests include signal processing and wireless communications.