Energy-Efficient Cell-Free Massive MIMO Through Sparse Large-Scale Fading Processing

Shuaifei Chen, *Graduate Student Member, IEEE*, Jiayi Zhang, *Senior Member, IEEE*, Emil Björnson, *Fellow, IEEE*, Özlem Tuğfe Demir, *Member, IEEE*, and Bo Ai, *Fellow, IEEE*

***Abstract*— Cell-free massive multiple-input multiple-output (CF mMIMO) systems serve the user equipments (UEs) by geographically distributed access points (APs) by means of joint transmission and reception. To limit the power consumption due to fronthaul signaling and processing, each UE should only be served by a subset of the APs, but it is hard to identify that subset. Previous works have tackled this combinatorial problem heuristically. In this paper, we propose a sparse distributed processing design for CF mMIMO, where the AP-UE association and long-term signal processing coefficients are jointly optimized. We formulate two sparsity-inducing mean-squared error (MSE) minimization problems and solve them by using efficient proximal approaches with block-coordinate descent. For the downlink, more specifically, we develop a virtually optimized large-scale fading precoding (V-LSFP) scheme using uplink-downlink duality. The numerical results show that the proposed sparse processing schemes work well in both uplink and downlink. In particular, they achieve almost the same spectral efficiency as if all APs would serve all UEs, while the energy efficiency is 2-4 times higher thanks to the reduced processing and signaling.**

***Index Terms*— Cell-free massive MIMO, energy efficiency, distributed processing, large-scale fading, sparse optimization.**

Manuscript received 14 August 2022; revised 9 February 2023; accepted 15 April 2023. Date of publication 1 May 2023; date of current version 12 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1807201, in part by the National Natural Science Foundation of China under Grant 61971027 and Grant 62221001, in part by the Beijing Natural Science Foundation under Grant L202013, and in part by the Natural Science Foundation of Jiangsu Province Major Project under Grant BK20212002. The work of Emil Björnson was supported by the Swedish Foundation for Strategic Research under Grant FFL18-0277. An earlier version of this paper was presented in part at the 2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication [DOI: 10.1109/SPAWC51304.2022.9834001]. The associate editor coordinating the review of this article and approving it for publication was X. Yuan. *(Corresponding author: Jiayi Zhang.)*

Shuaifei Chen is with the School of Electronic and Information Engineering and the Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China, and also with Purple Mountain Laboratories, Nanjing 211111, China (e-mail: shuaifeichen@bjtu.edu.cn).

Jiayi Zhang is with the School of Electronic and Information Engineering and the Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China (e-mail: jiayizhang@bjtu.edu.cn).

Emil Björnson is with the Department of Computer Science, KTH Royal Institute of Technology, 164 40 Kista, Sweden (e-mail: emilbjo@kth.se).

Özlem Tuğfe Demir was with the Department of Computer Science, KTH Royal Institute of Technology, 164 40 Kista, Sweden. She is now with the Department of Electrical and Electronics Engineering, TOBB University of Economics and Technology, 06560 Ankara, Turkey (e-mail: ozlemtugfedemir@etu.edu.tr).

Bo Ai is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China, and also with the Henan Joint International Research Laboratory of Intelligent Networking and Data Analysis, Zhengzhou University, Zhengzhou 450001, China (e-mail: boai@bjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3270299.

Digital Object Identifier 10.1109/TWC.2023.3270299

I. Introduction

As THE number of active wireless devices is steadily growing [2], the increasing requirements and demands for wireless communications force academia and industry to consider not only "how much and fast" the information can be transferred but also "how green" the networks can become in terms of the energy efficiency (EE). This shift in perception makes EE as important as a performance metric as spectral efficiency (SE) for fifth-generation (5G) networks [3]. During data transmission, the EE is defined as the ratio between the data rate and total power consumption [4]. Cellular massive multiple-input multiple-output (mMIMO) with access points (APs) equipped with large antenna arrays became the key technology for simultaneously improving the SE and EE in 5G [5], [6], [7], [8]. Looking towards the future, the main limiting factors for the SE and EE have now become the inter-cell interference caused by lack of cooperation between the APs, the large pathlosses between the APs and the user equipments (UEs) when using a small number of elevated APs, and the internal hardware energy consumption of the APs themselves [9]. The sixth-generation (6G) networks are expected to improve the SE and EE gains by $100\times$ over 5G networks [10] and must address these issues. This requires a denser network infrastructure operating in a cell-free (CF) manner that shifts the network from cell-centric to user-centric, and thus, provides ubiquitous coverage, improved network SE, and improved EE [11], [12], [13].

In the past few years, user-centric CF mMIMO has attracted extensive attention from the research community [14]. This paradigm inherits the *interference suppression gain* enabled by multiple antennas per AP from Cellular mMIMO and improves the *macro-diversity gain* by increasing the AP deployment density. In CF mMIMO systems, a large number of distributed APs are collaborating through a central processing unit (CPU) to serve the UEs with coherent joint transmission and reception, as illustrated in Fig. 1. This increases the average and worst-case data rates and reduces the total power consumption. Thus, CF mMIMO is envisioned as a promising paradigm shift for 6G networks [10]. The key difference from previous

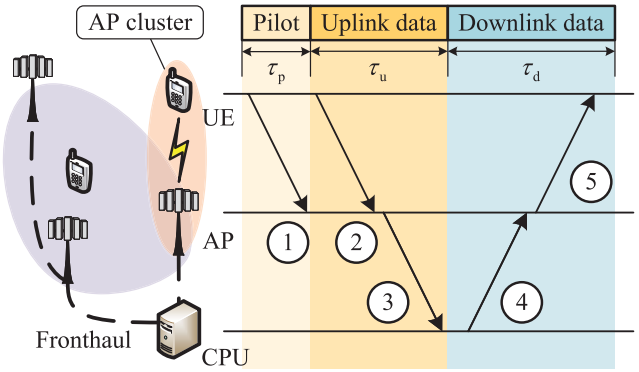


Fig. 1. Each UE is served by a subset of the APs in our considered user-centric CF mMIMO system. With the distributed operation, the signal processing tasks are divided between the APs and the CPU as indicated for 1) channel estimation, 2) local receive combining, 3) data decoding, 4) data encoding, and 5) local transmit precoding.

coordinated multipoint approaches is the dense deployment, user-centric approach, and signal processing schemes inherited from Cellular mMIMO.

Due to the UE-AP-CPU architecture, the signal processing tasks in CF mMIMO systems can be distributed between the APs and the CPU in different ways [15], [16]. According to how many of the tasks are delegated to the APs, CF mMIMO can operate in a *centralized* or *distributed* manner. In the centralized operation, the APs act as relays between the UEs and the CPU, which performs channel estimation and all signal processing by exploiting the instantaneous channel state information (CSI) gathered from the APs via the fronthaul connections. Although the centralized operation exhibits higher *user-experienced* data rates (i.e., 95%-likely SE) than its distributed alternative, it requires much higher computational complexity. As illustrated in Fig. 1, the alternative distributed operation is a two-stage processing procedure, in which *each* AP locally performs channel estimation and signal processing based on those estimates, while the CPU is only responsible for the final or initial processing of data using scaling factors that only depend on the large-scale fading (LSF) coefficients. This two-stage technique of processing was originally proposed for Cellular mMIMO, where the central unit linearly combines messages from/to each AP corresponding to the UEs from different cells to effectively eliminate inter-cell interference. This is referred to as *LSF decoding (LSFD)* [17] for the uplink and as *LSF precoding (LSFP)* [18], [19] for the downlink. For CF mMIMO, early papers on the topic of distributed uplink operation proposed to simply take the average of the local messages from different APs at the CPU. This can perform poorly since it neglects the inter-AP LSF information that is also available at the CPU and some APs can do more harm than good when serving far-away UEs. With this consideration, the authors in [15] developed LSFD for uplink CF mMIMO. When it comes to the downlink, CF mMIMO *inherently* performs LSFP since the transmitted messages for different APs are all encoded at the CPU but scaled differently by the APs when doing power allocation. Therefore, the concept of the LSFP was not mentioned in the existing CF mMIMO literature. To demonstrate the connections, the terminology "LSFP" is anyway used to represent the two-stage downlink transmit power allocation.

Although the distributed operation of CF mMIMO achieves a good compromise between data rates and computational complexity compared to the fully centralized operation [14], it might not be energy efficient in its original form where *all APs serve all UEs* [15], [16]. It is unnecessary for an AP to waste its power, computational, and fronthaul resources to serve distant UEs (with weak channels) when those UEs have better channels to other APs [20]. The geometry induces a sparse structure on the practically meaningful AP-UE associations. Prior works have suggested associating each UE with a subset of APs in advance and then excluding APs not associated with this UE when computing the LSFD vector [21], [22], [23]. Since the problem is combinatorial, to our best of knowledge, only heuristic methods have been proposed; see [14] for a recent survey. However, treating the AP-UE association as a separate combinatorial problem from the LSFD design, which is employed to maximize the SE [15], is suboptimal. This motivates us to consider the association as a part of the uplink LSFD and downlink LSFP design and employ of sparsity-inducing methods to jointly solve the association problem and signal processing design.

Sparse optimization methods have many successful applications in the fields of signal processing, image processing, and computer vision [24]. Specific to wireless communications, sparse optimization has been applied for random access [25], activity detection [26], and node sleeping [27]. For example, the authors in [27] shut down some "unnecessary" APs in a CF mMIMO system while satisfying the requested SEs by formulating the sparse reconstruction problem as a mixed-integer second-order cone program, where the globally optimal solution is found by utilizing the branch-and-bound approach. Similarly, in [13], mixed-binary programming is exploited to activate only the minimal subsets of APs for each UE to reduce the end-to-end network power consumption, where CF mMIMO is implemented on top of a virtualized cloud radio access network.

A. Main Contributions

We develop an energy-efficient distributed processing framework for CF mMIMO systems, which makes use of sparsity methods but in a novel way. We formulate a new sparse optimization problem for CF mMIMO that minimizes the data mean-squared-error (MSE), to enforce sparsity on the LSFD and LSFP coefficients. In consequence, the data rates are barely deteriorated while the power consumption needed to achieve it is minimized, which leads to higher EE. Our major contributions are listed as follows:

- We propose the sparse LSF processing design for both uplink and downlink, where joint AP-UE association and LSFD/LSFP is achieved by formulating sparsity-inducing MSE-minimizing problems to push small LSFD/LSFP coefficients to zero. We consider two kinds of sparsity: element-wise (EW) and group-wise (GW).
- We solve these formulated sparsity problems efficiently by developing proximal algorithms with block-coordinate

descent (BCD). The proposed algorithms contain closed-form updates and, thus, operate faster than the well-used optimization tool CVX [28].

- We develop a novel virtually optimized LSFP (V-LSFP) scheme for downlink power allocation in CF mMIMO systems by using the uplink-downlink duality. It is interesting to achieve $1.7\times$ 95%-likely downlink SE compared to the benchmark using distributed fractional power allocation (FPA) [21], [29].
- We compare the proposed sparse schemes with their fully-connected alternatives (where all APs serve all UEs) [15], and their partial alternatives [14], [22] with the separate AP-UE association as in [21]. The simulation results show that the proposed sparse LSF schemes significantly improve the EE with only a slight SE loss compared to the benchmarks.

The conference version of this paper, [1], only considers the sparse LSFD (S-LSFD) design in the uplink with GW sparsity. Herein, we extend [1] to a more generalized case considering both uplink and downlink with EW and GW sparsities.

B. Paper Outline and Notation

The remainder of this paper is organized as follows. Section II introduces the system model for our considered CF mMIMO system. Section III elaborates on the distributed uplink transmissions with LSFD. The sparse processing with sparse optimization is developed in Section IV by formulating two sparsity-inducing problems. Section V extends the analysis and design to the downlink where the LSFP and corresponding sparse processing are proposed. In Section VI, the details of the power consumption model are provided along with the definition of EE. Section VII numerically evaluates the proposed schemes and compares them with the considered benchmarks. Finally, we draw the conclusions and implications in Section VIII.

1) Reproducible Research: The simulation results can be reproduced using the Matlab code and data files available at: https://github.com/ShuaifeiChen273/sparse-LSFprocess-CFmMIMO.

2) Notation: Boldface lowercase letters, $\mathbf{x}$, denote column vectors, boldface uppercase letters, $\mathbf{X}$, denote matrices, and calligraphic uppercase letters, $\mathcal{A}$, denote sets. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The superscripts $^{\mathrm{T}}$, *, and $^{\mathrm{H}}$ denote the transpose, conjugate, and conjugate transpose, respectively. $x_i = [\mathbf{x}]_i$, $(x)_+ = \max(x, 0)$, and $\mathrm{sign}(\cdot)$ is the signum function. $\mathbb{E}\{\cdot\}$ computes the expected values and $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$ denotes the multi-variate circularly symmetric complex Gaussian distribution with correlation matrix $\mathbf{R}$.

II. CF mMIMO System Model

We consider a CF mMIMO system that consists of K single-antenna UEs and L geographically distributed APs, each equipped with N antennas. We adopt the user-centric CF architecture, where each UE is served by a subset of the APs, as illustrated in Fig. 1. The AP subsets of different UEs may overlap and are selected based on the UEs' channel qualities and service requirements. We will optimize these subsets but for now, we denote by $\mathcal{D}_l \subset \{1, \ldots, K\}$ the subset of UEs served by AP l and denote by $\mathcal{M}_k \subset \{1, \ldots, L\}$ the subset of APs serving UE k. All APs are connected via fronthaul connections to a CPU, which is coordinating the signal processing of all UEs, while the actual processing is distributed over the APs.

We adopt the standard time division duplex (TDD) operation and block fading model, where the time-frequency resources are divided into coherence blocks so that the channel coefficients can be assumed fixed in each block. We consider spatially correlated Rayleigh fading, which implies that the channel between AP l and UE k denoted by $\mathbf{h}_{kl} \in \mathbb{C}^N$ takes an independent realization in each coherence block according to

$$\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_{kl}), \tag{1}$$

where $\mathbf{R}_{kl} \in \mathbb{C}^{N \times N}$ is the spatial correlation matrix and $\beta_{kl} \triangleq \mathrm{tr}(\mathbf{R}_{kl})/N$ is the LSF coefficient describing pathloss and shadowing. It is assumed that AP l knows the correlation matrices $\{\mathbf{R}_{kl} : \forall k\}$ of all UEs since these represent the long-term channel statistics [4].

Each coherence block is used for both uplink and downlink payload data transmission and some portion is also used for uplink pilots. More precisely, each coherence block of τ_{c} channel uses is divided into three phases: a) τ_{p} channel uses are dedicated for pilot transmission and channel estimation; b) τ_{u} channel uses for uplink payload data; and c) the remaining $\tau_{\mathrm{d}} = \tau_{\mathrm{c}} - \tau_{\mathrm{p}} - \tau_{\mathrm{u}}$ channel uses for downlink payload data. We adopt the two-stage distributed processing approach in this paper [14] (as illustrated in Fig. 1), where only the data decoding and encoding are delegated to the CPU. The other signal processing tasks are done at the APs.

A. Uplink Pilot Transmission and Channel Estimation

During the channel estimation, each AP locally estimates the channels based on the uplink pilot transmission from the UEs. We consider a mutually orthogonal set of τ_{p} pilot sequences that must be shared between the UEs because in practical large networks, we will likely have $\tau_{\mathrm{p}} \ll K$. We denote by t_k the index of the pilot assigned to UE k and by $\mathcal{S}_{t_k}$ the set of UEs sharing pilot t_k. When the UEs in $\mathcal{S}_{t_k}$ transmit pilot t_k, the received signal $\mathbf{y}^{\mathrm{p}}_{t_k l} \in \mathbb{C}^N$ at AP l (after taking the inner product of the received signal and the pilot sequence t_k) is [4, Sec. 3]

$$\mathbf{y}^{\mathrm{p}}_{t_k l} = \sum_{i \in \mathcal{S}_{t_k}} \sqrt{\tau_{\mathrm{p}} p_{\mathrm{p}}} \mathbf{h}_{il} + \mathbf{n}_{t_k l}, \tag{2}$$

where $\mathbf{n}_{t_k l} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the receiver noise with noise power σ^2 and p_{p} is the pilot transmit power of each UE. The *minimum MSE (MMSE)* estimate of $\mathbf{h}_{kl}$ is [4, Sec. 3]

$$\widehat{\mathbf{h}}_{kl} = \sqrt{\tau_{\mathrm{p}} p_{\mathrm{p}}} \mathbf{R}_{kl} \boldsymbol{\Psi}^{-1}_{t_k l} \mathbf{y}^{\mathrm{p}}_{t_k l} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{B}_{kl}), \tag{3}$$

where $\boldsymbol{\Psi}_{t_k l} = \mathbb{E}\{\mathbf{y}^{\mathrm{p}}_{t_k l} (\mathbf{y}^{\mathrm{p}}_{t_k l})^{\mathrm{H}}\} = \sum_{i \in \mathcal{S}_{t_k}} \tau_{\mathrm{p}} p_{\mathrm{p}} \mathbf{R}_{il} + \sigma^2 \mathbf{I}_N$ is the correlation matrix of $\mathbf{y}^{\mathrm{p}}_{t_k l}$ in (2) and $\mathbf{B}_{kl} = \tau_{\mathrm{p}} p_{\mathrm{p}} \mathbf{R}_{kl} \boldsymbol{\Psi}^{-1}_{t_k l} \mathbf{R}_{kl}$.

III. Uplink Data Transmissions With LSFD

In this section, we provide the details of the distributed implementation of uplink reception, which are needed to formulate our design problem. Each AP locally employs an arbitrary receive combining scheme to obtain local soft estimates of the UE data. These estimates are then gathered at the CPU, which combines them using the LSFD approach.

In the uplink data phase, the received signal $\mathbf{y}_l^{\mathrm{ul}} \in \mathbb{C}^N$ at AP l is a superposition of the signals from all UEs:

$$\mathbf{y}_l^{\mathrm{ul}} = \sum_{i=1}^{K} \mathbf{h}_{il} s_i + \mathbf{n}_l, \tag{4}$$

where $s_i \in \mathbb{C}$ is the signal transmitted by UE i, $p_i = \mathbb{E}\{|s_i|^2\}$ is the corresponding transmit power, and $\mathbf{n}_l \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the independent additive receiver noise. AP l selects the normalized local combining vector $\mathbf{v}_{kl} = \bar{\mathbf{v}}_{kl} / \sqrt{\mathbb{E}\{\|\bar{\mathbf{v}}_{kl}\|_2^2\}} \in \mathbb{C}^N$ for UE k and then computes its local estimate of s_k as

$$\widehat{s}_{kl} = \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{y}_l^{\mathrm{ul}}. \tag{5}$$

One good option is to use the *local MMSE (L-MMSE)* combining scheme [21]

$$\bar{\mathbf{v}}_{kl} = p_k \left(\sum_{i=1}^{K} p_i \left(\widehat{\mathbf{h}}_{il} \widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{R}_{il} - \mathbf{B}_{il} \right) + \sigma^2 \mathbf{I}_N \right)^{-1} \widehat{\mathbf{h}}_{kl} \tag{6}$$

that suppresses interference and minimizes the local MSE $\mathbb{E}\{|s_k - \widehat{s}_{kl}|^2 | \{\widehat{\mathbf{h}}_{il} : \forall i\}\}$. Alternatively, the maximum ratio (MR) processing scheme with $\bar{\mathbf{v}}_{kl} = \widehat{\mathbf{h}}_{kl}$ can be used. Note that, for a generic UE k, although $\mathbf{v}_{kl} \neq \mathbf{0}$ for all APs, only the serving APs in $\mathcal{M}_k$ need to compute $\mathbf{v}_{kl}$.

Next, the APs transfer their local data estimates to the CPU, which performs the final decoding of s_k by linearly combining the local estimates:

$$\widehat{s}_k = \sum_{l=1}^{L} a_{kl}^{\star} \widehat{s}_{kl} = \sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{y}_l^{\mathrm{ul}}, \tag{7}$$

where $a_{kl} \in \mathbb{C}$ is the weight that the CPU assigns to the local signal estimate $\widehat{s}_{kl}$. In LSFD, the CPU selects the weights $\{a_{kl}\}$ as a deterministic function of the channel statistics (to avoid sharing channel estimates [14]). Note that only those APs assigning a non-zero value to a_{kl} participate in the decoding, thus this formulation supports a user-centric architecture. For a given set $\{a_{kl}\}$ of LSFD weights, the serving APs of UE k can be extracted as $\mathcal{M}_k = \{l : a_{kl} \neq 0\}$.

By letting $\mathbf{g}_{ki} = [\mathbf{v}_{k1}^{\mathrm{H}} \mathbf{h}_{i1}, \ldots, \mathbf{v}_{kL}^{\mathrm{H}} \mathbf{h}_{iL}]^{\mathrm{T}} \in \mathbb{C}^L$ denote the receive-combined channels from UE i when receiving signals from UE k, and $\mathbf{a}_k = [a_{k1}, \ldots, a_{kL}]^{\mathrm{T}} \in \mathbb{C}^L$ denote the LSFD weight vector of UE k, the estimate of s_k in (7) can be rewritten as

$$\widehat{s}_k = \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk} s_k + \sum_{i=1, i \neq k}^{K} \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} s_i + n_k', \tag{8}$$

where $n_k' = \sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l$ is the resulting noise. The effective uplink channel $\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}$ in (8) is not known at the CPU but its average $\mathbb{E}\{\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}\} = \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}$ is deterministic and non-zero if the receive combiner is selected as suggested above. Therefore, it can be assumed to be available at the CPU and we can therefore quantify the achievable uplink SE using the *hardening bound* [4, Thm. 4.4]. More precisely, the resulting SE of UE k is

$$\mathsf{SE}_k^{\mathrm{ul}} = \frac{\tau_{\mathrm{u}}}{\tau_{\mathrm{c}}} \log_2 \left(1 + \mathsf{SINR}_k^{\mathrm{ul}} \right) \quad \text{bit/s/Hz}, \tag{9}$$

where

$$\mathsf{SINR}_k^{\mathrm{ul}} = \frac{|\mathbf{a}_k^{\mathrm{H}} \boldsymbol{\xi}_k|^2}{\mathbf{a}_k^{\mathrm{H}} \boldsymbol{\Delta}_k \mathbf{a}_k - |\mathbf{a}_k^{\mathrm{H}} \boldsymbol{\xi}_k|^2} = \frac{|\mathbf{a}_k^{\mathrm{H}} \boldsymbol{\xi}_k|^2}{\mathbf{a}_k^{\mathrm{H}} (\boldsymbol{\Delta}_k - \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{\mathrm{H}}) \mathbf{a}_k}, \tag{10}$$

is the effective uplink signal-to-interference-plus-noise ratio (SINR) [14, Thm. 5.4] with

$$\boldsymbol{\Delta}_k = \sum_{i=1}^{K} p_i \mathbb{E}\{\mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}}\} + \sigma^2 \mathbf{I}_L \in \mathbb{C}^{L \times L}, \tag{11}$$

$$\boldsymbol{\xi}_k = \sqrt{p_k} \mathbb{E}\{\mathbf{g}_{kk}\} \in \mathbb{C}^L. \tag{12}$$

We note that the effective uplink SINR in (10) is a generalized Rayleigh quotient with respect to $\mathbf{a}_k$. Hence, with the help of the generalized eigenvector result [4, Lem. B.10] and matrix inversion lemma [4, Lem. B.4], the optimal LSFD (O-LSFD) weight vector is

$$\mathbf{a}_k^{\mathrm{opt}} = c_k \boldsymbol{\Delta}_k^{-1} \boldsymbol{\xi}_k \tag{13}$$

with $c_k \in \mathbb{C}$ being an arbitrary non-zero scaling factor. The resulting maximum SINR value is $\mathsf{SINR}_k^{\mathrm{ul}} = \boldsymbol{\xi}_k^{\mathrm{H}} (\boldsymbol{\Delta}_k - \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{\mathrm{H}})^{-1} \boldsymbol{\xi}_k$.

Moreover, we notice that the uplink MSE in the data decoding of UE k is

$$\mathsf{MSE}_k^{\mathrm{ul}} = \mathbb{E}\{|s_k - \widehat{s}_k|^2\} = \mathbf{a}_k^{\mathrm{H}} \boldsymbol{\Delta}_k \mathbf{a}_k - 2\sqrt{p_k} \Re(\mathbf{a}_k^{\mathrm{H}} \boldsymbol{\xi}_k) + p_k, \tag{14}$$

which is minimized by the LSFD vector

$$\mathbf{a}_k^{\mathrm{mse}} = \sqrt{p_k} \boldsymbol{\Delta}_k^{-1} \boldsymbol{\xi}_k, \tag{15}$$

which is equal to $\mathbf{a}_k^{\mathrm{opt}}$ in (13) if the scaling factor is set to $c_k = \sqrt{p_k}$. While there is only one LSFD vector minimizing the MSE, we can use any scaling factor to maximizing the SINR. We conclude that we can identify an optimal LSFD vector by minimizing the MSE instead of maximizing the SINR, which is a feature that we will exploit in the remainder of this paper.

By using the notation $\mathbf{a} = [\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{KL}$, $\boldsymbol{\xi} = [\sqrt{p_1} \boldsymbol{\xi}_1^{\mathrm{T}}, \ldots, \sqrt{p_K} \boldsymbol{\xi}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{KL}$, and $\boldsymbol{\Delta} = \mathrm{diag}(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K) \in \mathbb{C}^{KL \times KL}$, we can express the uplink sum MSE of all UEs as

$$\sum_{k=1}^{K} \mathsf{MSE}_k^{\mathrm{ul}} = \mathbf{a}^{\mathrm{H}} \boldsymbol{\Delta} \mathbf{a} - 2\Re(\mathbf{a}^{\mathrm{H}} \boldsymbol{\xi}) + \sum_{k=1}^{K} p_k. \tag{16}$$

Recall that in (14), each uplink MSE only depends on the respective UE's LSFD vector $\mathbf{a}_k$. Hence, finding the collective LSFD vector $\mathbf{a}^{\mathrm{opt}}$ that minimizes the sum MSE $\sum_{k=1}^{K} \mathsf{MSE}_k^{\mathrm{ul}}$ is equivalent to finding the set of O-LSFD vectors $\{\mathbf{a}_k^{\mathrm{opt}} : k = 1, \ldots, K\}$ that simultaneously minimize their corresponding uplink MSEs.

IV. Sparse LSFD With MSE Minimization

One way to implicitly obtain the AP selection for UE k is to first design a suitable LSFD vector as if all APs serve the UE and then let only the APs with non-zero weights serve it: $\mathcal{M}_k = \{l : a_{kl} \neq 0,\ l = 1, \ldots, L\}$. The problem with this approach is that the O-LSFD vector $\mathbf{a}^{\text{opt}}$ in (13) in general only contains non-zero values, so all APs would have to serve all UEs. However, we have noticed that $\mathbf{a}^{\text{opt}}$ typically contains a few large values and many small values due to the natural pathloss differences between APs and UEs in a distributed deployment. In this section, we will propose the S-LSFD design that resembles O-LSFD but pushes small weights to zero, thereby greatly limiting how many APs must serve each UE.

A. Problem Formulation

We recall that O-LSFD is obtained by minimizing the quadratic form in (16) with respect to the collective LSFD vector $\mathbf{a}$. Inspired by this fact and sparse reconstruction methods, we propose the generic real-valued MSE minimization problem

$$\min_{\underline{\mathbf{a}} \in \mathbb{R}^{2KL}} \underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\Delta}} \underline{\mathbf{a}} - 2\underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\xi}} + \Omega(\underline{\mathbf{a}}) \tag{17}$$

with the real variables $\underline{\mathbf{a}} = [\underline{\mathbf{a}}_1^{\mathrm{T}}, \ldots, \underline{\mathbf{a}}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2KL}$, $\underline{\boldsymbol{\xi}} = [\sqrt{p_1}\underline{\boldsymbol{\xi}}_1^{\mathrm{T}}, \ldots, \sqrt{p_K}\underline{\boldsymbol{\xi}}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2KL}$, and $\underline{\boldsymbol{\Delta}} = \mathrm{diag}(\underline{\boldsymbol{\Delta}}_1, \ldots, \underline{\boldsymbol{\Delta}}_K) \in \mathbb{R}^{2KL \times 2KL}$, where

$$\underline{\mathbf{a}}_k = \begin{bmatrix} \Re(\mathbf{a}_k) \\ \Im(\mathbf{a}_k) \end{bmatrix} \in \mathbb{R}^{2L},\ \underline{\boldsymbol{\xi}}_k = \begin{bmatrix} \Re(\boldsymbol{\xi}_k) \\ \Im(\boldsymbol{\xi}_k) \end{bmatrix} \in \mathbb{R}^{2L},$$
$$\underline{\boldsymbol{\Delta}}_k = \begin{bmatrix} \Re(\boldsymbol{\Delta}_k) & -\Im(\boldsymbol{\Delta}_k) \\ \Im(\boldsymbol{\Delta}_k) & \Re(\boldsymbol{\Delta}_k) \end{bmatrix} \in \mathbb{R}^{2L \times 2L}. \tag{18}$$

The first two terms in (17) $\underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\Delta}} \underline{\mathbf{a}} - 2\underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\xi}}$ represent the "MSE" cost, which is a convex function of $\underline{\mathbf{a}}$. The third term $\Omega(\underline{\mathbf{a}})$ is a sparsity-inducing function that can be designed to encourage small values in $\underline{\mathbf{a}}$ to become zero at the optimal solution. We refer to the minimizer of (17) as a S-LSFD vector. By selecting different $\Omega(\underline{\mathbf{a}})$, different sparsity patterns can be achieved in the LSFD vector $\underline{\mathbf{a}}$. We will consider two key examples in this section.

B. EW Sparsity and Proximal Algorithm

Element-wise (EW) sparsity can be induced on the LSFD vector by using the ℓ_1-norm, as

$$\Omega(\underline{\mathbf{a}}) = \lambda \|\underline{\mathbf{a}}\|_1, \tag{19}$$

where $\lambda \geq 0$ is a tunable EW regularization parameter. The physical interpretation behind EW sparsity is to limit the average number of UEs that each AP serves, but otherwise letting the optimization problem select freely which AP-UE associations that should remain. A large value of λ induces more EW sparsity. When using (19), the optimization problem in (17) becomes

$$\mathsf{P}^{\text{ew}}: \quad \min_{\underline{\mathbf{a}} \in \mathbb{R}^{2KL}} \underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\Delta}} \underline{\mathbf{a}} - 2\underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\xi}} + \lambda \|\underline{\mathbf{a}}\|_1, \tag{20}$$

which is convex since the ℓ_1-norm penalty is a convex function. In fact, we can solve the K subproblems

$$\mathsf{P}_k^{\text{ew}}: \quad \min_{\underline{\mathbf{a}}_k \in \mathbb{R}^{2L}} f(\underline{\mathbf{a}}_k) + \lambda \|\underline{\mathbf{a}}_k\|_1,\ k = 1, \ldots, K \tag{21}$$

in parallel to obtain the solution to (20) since both the "MSE" cost and sparsity function can be decoupled between the UEs, where $f(\underline{\mathbf{a}}_k) = \underline{\mathbf{a}}_k^{\mathrm{T}} \underline{\boldsymbol{\Delta}}_k \underline{\mathbf{a}}_k - 2\sqrt{p_k}\underline{\mathbf{a}}_k^{\mathrm{T}} \underline{\boldsymbol{\xi}}_k$.

Since the subproblems $\{\mathsf{P}_k^{\text{ew}}\}$ in (21) are convex with non-smooth sparsity-inducing penalties, the proximal methods can be utilized to solve them efficiently [30].

By using the proximal methods, we start with an initial point $\underline{\mathbf{a}}_k^0$ which can be initialized by its corresponding O-LSFD vector $\mathbf{a}_k^{\text{opt}}$ using (18), and then compute a sequence of updates $\underline{\mathbf{a}}_k^n$ that converges to the optimal solution to (21), where n is the iteration index. Given the $\underline{\mathbf{a}}_k^n$ obtained at iteration n, we can find the next update $\underline{\mathbf{a}}_k^{n+1}$ by solving the following *proximal problem*

$$\min_{\underline{\mathbf{a}}_k \in \mathbb{R}^{2L}} \frac{1}{2} \|\underline{\mathbf{a}}_k - G(\underline{\mathbf{a}}_k^n)\|_2^2 + \mu\lambda \|\underline{\mathbf{a}}_k\|_1, \tag{22}$$

where $G(\underline{\mathbf{a}}_k^n) = \underline{\mathbf{a}}_k^n - \mu \nabla f(\underline{\mathbf{a}}_k^n)$ is the so-called *gradient update* and μ is the step length which can be computed in practice via line search [30]. The unique solution of (22) can be found due to the strong convexity [30]. This is given as follows.

Lemma 1: Since $\nabla f(\underline{\mathbf{a}}_k) = 2\underline{\boldsymbol{\Delta}}_k \underline{\mathbf{a}}_k - 2\sqrt{p_k}\underline{\boldsymbol{\xi}}_k$, the unique solution of (22) can be obtained as

$$\mathrm{Prox}_{\mu\lambda,\ell_1}(G(\underline{\boldsymbol{\alpha}}_k^n)) = \arg\min_{\underline{\mathbf{a}}_k \in \mathbb{R}^{2L}} \frac{1}{2} \|\underline{\mathbf{a}}_k - G(\underline{\mathbf{a}}_k^n)\|_2^2 + \mu\lambda \|\underline{\mathbf{a}}_k\|_1, \tag{23}$$

which is the proximal operator of the ℓ_1-norm [30] and can be componentwisely computed as

$$[\mathrm{Prox}_{\mu,\ell_1}(\mathbf{u})]_i = \mathrm{sign}(u_i) \cdot (|u_i| - \mu)_+. \tag{24}$$

Proof: The proof follows the results in [30] and is omitted due to limited space. ∎

To obtain the minimizer of P_k^{ew} in (21), the vector can be updated as

$$\underline{\mathbf{a}}_k^{n+1} \leftarrow \mathrm{Prox}_{\mu\lambda,\ell_1}(G(\hat{\underline{\mathbf{a}}}_k^n)) \tag{25}$$

with the Nesterov step $\hat{\underline{\mathbf{a}}}_k^n = \underline{\mathbf{a}}_k^n + \frac{n-1}{n+2}(\underline{\mathbf{a}}_k^n - \underline{\mathbf{a}}_k^{n-1})$ that is known to accelerate the convergence to the solution of (21) [30]. By performing the inverse transformations in (18), we achieve the complex-valued EW S-LSFD vectors.

C. GW Sparsity and Proximal Algorithm With BCD

The EW sparsity approach limits the average number of UEs served by an AP, but without inducing any preference on how the UE load is distributed among the APs. In practice, we might prefer that some APs are not serving any UEs at all, so that we can save power by putting them into sleep mode. This property can be encouraged by also inducing group-wise (GW) sparsity on the LSFD vector. More precisely, we propose

to use the composite $\ell_1+\ell_1/\ell_2$-norm to simultaneously induce GW and EW sparsity:

$$\Omega(\underline{\mathbf{a}}) = \gamma \sum_{l=1}^{L} \|\boldsymbol{x}_l\|_2 + \lambda \|\underline{\mathbf{a}}\|_1,$$
$$\boldsymbol{x}_l = [\Re(\boldsymbol{a}_l)^{\mathrm{T}}, \Im(\boldsymbol{a}_l)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2K}, \tag{26}$$

where γ is the tunable GW regularization parameter and $\boldsymbol{a}_l = [a_{1l}, \ldots, a_{Kl}]^{\mathrm{T}} \in \mathbb{C}^K$ is subset of the vector $\mathbf{a}$ related to AP l. Larger value of γ induces more GW sparsity on vector $\underline{\mathbf{a}}$. The first term in (26) is a ℓ_1/ℓ_2-norm that behaves like a ℓ_1-norm applied to the vector $[\|\boldsymbol{x}_1\|_2, \ldots, \|\boldsymbol{x}_L\|_2]^{\mathrm{T}}$. Element l, i.e., $\|\boldsymbol{x}_l\|_2$, is small if AP l has little impact on the decoding and thus the ℓ_1/ℓ_2-norm promotes making such values identically zero (i.e., inactivate the AP). The second term limits the number of UEs served by the remaining active APs. The sparse problem in (17) becomes

$$\mathsf{P}^{\mathrm{gw}}: \quad \min_{\underline{\mathbf{a}} \in \mathbb{R}^{2KL}} \underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\Delta}} \underline{\mathbf{a}} - 2\underline{\mathbf{a}}^{\mathrm{T}} \underline{\boldsymbol{\xi}} + \gamma \sum_{l=1}^{L} \|\boldsymbol{x}_l\|_2 + \lambda \|\underline{\mathbf{a}}\|_1 \tag{27}$$

which is convex since the composite $\ell_1 + \ell_1/\ell_2$-norm penalty is a convex function.

The ℓ_1/ℓ_2-norm term restricts (27) from being decomposed into K subproblems that can be solved in parallel, in contrast to (20) in the EW case. But fortunately, (27) is separable between the APs so that the BCD approach can be used to guarantee convergence to the global optimum of (27) [30]. We equivalently rewrite the original problem P^{gw} as

$$\min_{\underline{\mathbf{a}} \in \mathbb{R}^{2KL}} \left\| \bar{\boldsymbol{\xi}} - \sum_{l=1}^{L} \mathbf{X}_l \boldsymbol{x}_l \right\|_2^2 + \gamma \sum_{l=1}^{L} \|\boldsymbol{x}_l\|_2 + \lambda \|\underline{\mathbf{a}}\|_1, \tag{28}$$

where $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \underline{\boldsymbol{\Delta}}$ and $\mathbf{X}_l \in \mathbb{R}^{2KL \times 2K}$ is the submatrix of $\mathbf{X}$ with columns corresponding to group l such that $\mathbf{X}\underline{\mathbf{a}} = \sum_{l=1}^{L} \mathbf{X}_l \boldsymbol{x}_l$. We use the notation $\bar{\boldsymbol{\xi}} = (\mathbf{X}^{\mathrm{T}})^{-1}\underline{\boldsymbol{\xi}} \in \mathbb{R}^{2KL}$. Note that (28) has the same form as the so-called "sparse-group Lasso" problem [31]. Hence, by using the BCD approach, we can solve (28) efficiently by iteratively minimizing the subproblem of group l while fixing the coefficients of the other groups:

$$\mathsf{P}_l^{\mathrm{gw}}: \quad \min_{\boldsymbol{x}_l \in \mathbb{R}^{2K}} g(\boldsymbol{x}_l) + \Omega'(\boldsymbol{x}_l), \quad l = 1, \ldots, L, \tag{29}$$

where $g(\boldsymbol{x}_l) = \|\mathbf{r}_l - \mathbf{X}_l \boldsymbol{x}_l\|_2^2$, $\Omega'(\boldsymbol{x}_l) = \gamma\|\boldsymbol{x}_l\|_2 + \lambda\|\boldsymbol{x}_l\|_1$, and $\mathbf{r}_l = \bar{\boldsymbol{\xi}} - \sum_{j \neq l} \mathbf{X}_j \boldsymbol{x}_j$ is the partial residual of $\bar{\boldsymbol{\xi}}$ subtracting all group coefficients except group l. $\Omega'(\boldsymbol{x}_l)$ implies that for group l, the other group coefficients are considered fixed and their penalties can be ignored.

Similar to solving $\mathsf{P}_k^{\mathrm{ew}}$, for a subproblem $\mathsf{P}_l^{\mathrm{gw}}$ of group l, given the current $\boldsymbol{x}_l^n$ obtained at iteration n, the next update $\boldsymbol{x}_l^{n+1}$ is found by solving the following proximal problem

$$\min_{\boldsymbol{x}_l \in \mathbb{R}^{2K}} \frac{1}{2} \|\boldsymbol{x}_l - G(\boldsymbol{x}_l^n)\|_2^2 + \mu \Omega'(\boldsymbol{x}_l), \tag{30}$$

where $G(\boldsymbol{x}_l^n) = \boldsymbol{x}_l^n - \mu \nabla g(\boldsymbol{x}_l^n)$. The unique solution to (30) can be found due to the strong convexity [30] and is given as follows.

Algorithm 1 Algorithm for Warm-Restart

Input: λ, $\bar{\lambda}$, $\eta \in (0, 1)$
Output: $\mathbf{a} \in \mathbb{C}^{KL}$

```
λ' = λ̄;
(Outer loop) while λ' ≥ λ do
    (Inner loop) Solve the considered sparsity problem
      with λ' and update a;
    if λ' = λ then
        Break;
    else
        λ' ← max(ηλ', λ);
```

Lemma 2: Since $\nabla g(\boldsymbol{x}_l) = 2\mathbf{X}_l^{\mathrm{T}}(\mathbf{X}_l \boldsymbol{x}_l - \mathbf{r}_l)$, *the unique solution of (30) can be computed as*

$$\mathrm{Prox}_{\mu,\Omega'}(G(\boldsymbol{x}_l^n)) = \mathrm{Prox}_{\mu\gamma,\ell_2} \circ \mathrm{Prox}_{\mu\lambda,\ell_1}(G(\boldsymbol{x}_l^n)), \tag{31}$$

where $f \circ g(x) \triangleq f(g(x))$ *for any function* f *and* g,

$$\mathrm{Prox}_{\mu,\ell_2}(\mathbf{u}) = \begin{cases} \dfrac{\mathbf{u}}{\|\mathbf{u}\|_2}(\|\mathbf{u}\|_2 - \mu)_+, & \text{if } \mathbf{u} \neq \mathbf{0}, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{32}$$

is the proximal operator of the ℓ_2*-norm [30], and* $\mathrm{Prox}_{\mu,\ell_1}(\mathbf{x})$ *is the proximal operator of the* ℓ_1*-norm given in (24).*

Proof: The proof follows a similar approach as in [31], but for problem (28). The details are given in Appendix A for completeness. ∎

The minimizer of group l can be updated as

$$\boldsymbol{x}_l^{n+1} \leftarrow \mathrm{Prox}_{\mu,\Omega'}(G(\hat{\boldsymbol{x}}_l^n)) \tag{33}$$

with the Nesterov step $\hat{\boldsymbol{x}}_l^n = \boldsymbol{x}_l^n + \frac{n-1}{n+2}(\boldsymbol{x}_l^n - \boldsymbol{x}_l^{n-1})$ accelerating the convergence [30], and is then fixed while the other groups are minimized until next iteration. By iteratively updating $\{\mathsf{P}_l^{\mathrm{gw}} : l = 1, \ldots, L\}$, the global solution to (29) can be reached. With the inverse transformations in (26), we achieve the complex-valued GW S-LSFD vectors. Although the mixed sparse penalty in (26) is more generalized than the penalty in (19), it requires more computational complexity.

D. Algorithm Implementation

We have noticed that the EW and GW sparsity problems are faster to solve with a larger regularization parameter λ. With this consideration in mind, we propose to perform the warm-restart strategy [30] on λ, which accelerates the convergence by solving a sequence of simple subproblems. The warm-restart strategy starts with a large regularization term $\bar{\lambda} \gg \lambda$, then iteratively shrinks $\bar{\lambda}$ towards λ and solves the corresponding subproblems. In each iteration, the subproblem is solved by employing the solution to the previous subproblem as the initialization. In other words, the warm-restart strategy used in our scenario operates as a sequence of nested loops, which is summarized in Algorithm 1. The considered sparse problems P^{ew} and P^{gw} can be solved by performing Algorithm 2 and Algorithm 3, respectively. These algorithms can be initialized by the O-LSFD vectors without sparsity and

Algorithm 2 Algorithm for Solving P^{ew}

Input: $\underline{\boldsymbol{\Delta}} \in \mathbb{R}^{2KL\times 2KL}$, $\underline{\mathbf{a}} \in \mathbb{R}^{2KL}$, $\underline{\boldsymbol{\xi}} \in \mathbb{R}^{2KL}$, λ, μ, $n_{\max}$
Output: $\mathbf{a} \in \mathbb{C}^{KL}$

for $k = 1, \ldots, K$ **do**
  $n = 1$;
  $\underline{\mathbf{a}}_k^- \leftarrow \underline{\mathbf{a}}_k$;
  repeat
    $\hat{\underline{\mathbf{a}}}_k = \underline{\mathbf{a}}_k + \frac{n-1}{n+2}(\underline{\mathbf{a}}_k - \underline{\mathbf{a}}_k^-)$;
    Compute $\mathrm{Prox}_{\mu\lambda,\ell_1}(G(\hat{\underline{\mathbf{a}}}_k))$ with the help of (24);
    $\underline{\mathbf{a}}_k \leftarrow \mathrm{Prox}_{\mu\lambda,\ell_1}(G(\hat{\underline{\mathbf{a}}}_k))$;
    $\underline{\mathbf{a}}_k^- \leftarrow \underline{\mathbf{a}}_k$;
    $n \leftarrow n + 1$;
  until $n = n_{\max}$ *or convergence*;
Obtain $\mathbf{a} \in \mathbb{C}^{KL}$ with the inverse transformation in (18).

terminated when the maximum number of iterations $n_{\max}$ is reached or convergence, measured by the change in objective function value.

V. Downlink Transmissions With LSFP

In this section, we consider the distributed downlink transmission with the goal of limiting the number of APs that serve each UE and the number of active APs. The downlink payload data of each UE is first sent to the APs that serve it. Next, the data symbols are locally precoded at the APs with local precoding vectors designed based on instantaneous channel estimates and then transmitted using AP-specific power coefficients. These coefficients are designed based on long-term statistics and, thus, correspond to LSFP in the Cellular literature [18]. We extend the sparse optimization to the downlink and develop a sparse LSFP (S-LSFP) design where the joint AP-UE association and LSFP is achieved.

In the downlink data phase, the distributed implementation is realized by constructing a linearly combined precoded signals from each AP. Let $\varsigma_i \in \mathbb{C}$ denote the unit-power downlink data signal intended for UE i with $\mathbb{E}\{|\varsigma_i|^2\} = 1$. The data signals for different UEs $\{\varsigma_i : i = 1, \ldots, K\}$ are independent. For a generic AP l, the CPU encodes the related symbols $\{\varsigma_i : i \in \mathcal{D}_l\}$ and transfers them to AP l via the fronthaul links. Then, AP l constructs the transmitted signal as

$$\mathbf{x}_l = \sum_{i=1}^{K} \sqrt{\rho_{il}}\mathbf{w}_{il}\varsigma_i, \tag{34}$$

where $\mathbf{w}_{il} = \bar{\mathbf{w}}_{il}/\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{il}\|_2^2\}} \in \mathbb{C}^N$ is the normalized precoding vector that AP l selects for UE i such that $\mathbb{E}\{\|\mathbf{w}_{il}\|_2^2\} = 1$. The precoding vector $\bar{\mathbf{w}}_{il}$ can have an arbitrary norm, while $\mathbf{w}_{il}$ has unit long-term power. Therefore, $\mathbf{w}_{il}$ only specifies the precoding direction, whereas the power allocation coefficient ρ_{il} controls the power. In the CF mMIMO literature, the precoding vectors

Algorithm 3 Algorithm for Solving P^{gw}

Input: $\underline{\boldsymbol{\Delta}} \in \mathbb{R}^{2KL\times 2KL}$, $\mathbf{a} \in \mathbb{C}^{KL}$, $\underline{\boldsymbol{\xi}} \in \mathbb{R}^{2KL}$, γ, λ, μ, $n_{\max}$
Output: $\mathbf{a} \in \mathbb{C}^{KL}$

Compute $\mathbf{X} \in \mathbb{R}^{2KL\times 2KL}$ such that $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \underline{\boldsymbol{\Delta}}$;
$\bar{\boldsymbol{\xi}} = (\mathbf{X}^{\mathrm{T}})^{-1}\underline{\boldsymbol{\xi}}$;
for $l = 1, \ldots, L$ **do**
  Extract $\boldsymbol{a}_l$ from $\mathbf{a}$ as $\boldsymbol{a}_l = [a_{1l}, \ldots, a_{Kl}]^{\mathrm{T}} \in \mathbb{C}^K$ such that $\boldsymbol{x}_l = [\Re(\boldsymbol{a}_l)^{\mathrm{T}}, \Im(\boldsymbol{a}_l)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2K}$ and extract the corresponding $\mathbf{X}_l \in \mathbb{R}^{2KL\times 2K}$ from $\mathbf{X}$;
  Compute the partial residual $\mathbf{r}_l = \bar{\boldsymbol{\xi}} - \sum_{j\neq l}\mathbf{X}_j\boldsymbol{x}_j \in \mathbb{R}^{2KL}$;
  $n = 1$;
  $\boldsymbol{x}_l^- \leftarrow \boldsymbol{x}_l$;
  repeat
    $\hat{\boldsymbol{x}}_l = \boldsymbol{x}_l + \frac{n-1}{n+2}(\boldsymbol{x}_l - \boldsymbol{x}_l^-)$;
    Compute $\mathrm{Prox}_{\mu,\Omega'}(G(\hat{\boldsymbol{x}}_l))$ with the help of (31), (24), and (32);
    $\boldsymbol{x}_l \leftarrow \mathrm{Prox}_{\mu,\Omega'}(G(\hat{\boldsymbol{x}}_l))$;
    $\boldsymbol{x}_l^- \leftarrow \boldsymbol{x}_l$;
    $n \leftarrow n + 1$;
  until $n = n_{\max}$ *or convergence*;
  Update $\mathbf{a}$ by replacing the elements indexed by l;
Obtain $\mathbf{a} \in \mathbb{C}^{KL}$ with the inverse transformation in (26).

$\{\mathbf{w}_{kl} : l = 1, \ldots, L, k = 1, \ldots, K\}$ are normally selected to match with the uplink combining vectors as

$$\mathbf{w}_{kl} = \mathbf{v}_{kl} = \frac{\bar{\mathbf{v}}_{kl}}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{v}}_{kl}\|^2\}}}. \tag{35}$$

This can be motivated by uplink-downlink duality [21] and we will derive a similar result below.

The received signal $y_k^{\mathrm{dl}} \in \mathbb{C}$ at UE k is

$$\begin{aligned} y_k^{\mathrm{dl}} &= \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}}\mathbf{x}_l + n_k \\ &= \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}}\sqrt{\rho_{kl}}\mathbf{w}_{kl}\varsigma_k + \sum_{i=1,i\neq k}^{K}\left(\sum_{l=1}^{L}\mathbf{h}_{kl}^{\mathrm{H}}\sqrt{\rho_{il}}\mathbf{w}_{il}\varsigma_i\right) \\ &\quad + n_k, \end{aligned} \tag{36}$$

where $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is the independent receiver noise.

Using the combining vectors from the uplink as in (35), $\mathbf{g}_{ik} = [\mathbf{w}_{i1}^{\mathrm{H}}\mathbf{h}_{k1}, \ldots, \mathbf{w}_{iL}^{\mathrm{H}}\mathbf{h}_{kL}]^{\mathrm{T}} \in \mathbb{C}^L$ represents the precoded channels to UE k when the APs transmit to UE i. We define the vector whose elements are the square roots of the power coefficients that the different APs assign to UE k as

$$\mathbf{b}_k = [\sqrt{\rho_{k1}}, \ldots, \sqrt{\rho_{kL}}]^{\mathrm{T}} = \sqrt{\rho_k}\boldsymbol{\omega}_k \in \mathbb{C}^L \tag{37}$$

as the *LSFP vector* of UE k, where ρ_k is the total transmit power for UE k and $\boldsymbol{\omega}_k = [\omega_{k1}, \ldots, \omega_{kL}]^{\mathrm{T}}$ is a unit-norm vector with non-negative entries indicating how the power is allocated among the APs, which is the main concern of this

paper. Notice that $\{\mathbf{b}_k : k = 1, \ldots, K\}$ can be optimized by the CPU in a network-wide manner to maximize certain utilities only employing channel statistics, which is why it is called LSFP.

We can rewrite the received signal at UE k in (36) as

$$y_k^{\mathrm{dl}} = \mathbf{g}_{kk}^{\mathrm{H}} \mathbf{b}_k \varsigma_k + \sum_{i=1, i \neq k}^{K} \mathbf{g}_{ik}^{\mathrm{H}} \mathbf{b}_i \varsigma_i + n_k, \tag{38}$$

where $\{\mathbf{g}_{ik}^{\mathrm{H}} \mathbf{b}_i : i = 1, \ldots, K\}$ represent the effective downlink channels. We can now compute an achievable downlink SE at UE k by utilizing the hardening bound [4, Thm. 4.6], as

$$\mathsf{SE}_k^{\mathrm{dl}} = \frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{c}}} \log_2 \left(1 + \mathsf{SINR}_k^{\mathrm{dl}}\right) \quad \text{bit/s/Hz} \tag{39}$$

where the effective downlink SINR is given by [14, Cor. 6.3]

$$\mathsf{SINR}_k^{\mathrm{dl}} = \frac{|\mathbf{b}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}|^2}{\sum_{i=1}^{K} \mathbb{E}\{|\mathbf{b}_i^{\mathrm{H}} \mathbf{g}_{ik}|^2\} - |\mathbf{b}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}|^2 + \sigma^2}. \tag{40}$$

Note that the SE holds for any local precoding and LSFP vectors. One important difference from LSFD in the uplink is that the SINR in (40) is not a generalized Rayleigh quotient with respect to the LSFP vectors. As can be seen from (40), the downlink SINR of a generic UE k is not only affected by the LSFP vector $\mathbf{b}_k$, but also by all other vectors, i.e., $\{\mathbf{b}_i : i = 1, \ldots, K\}$. Hence, it is not possible to obtain the optimal LSFP weights to maximize one UE's SE without affecting the others. The LSFP vectors should be optimized for a certain utility maximization and in general obtaining closed-form results is not possible.

To aid in identifying suitable LSFP vectors, we will establish a novel uplink-downlink duality between the LSFD and LSFP vectors by extending the approach in [21, Prop. 4] that considers the duality between centralized combining and precoding vectors based on the channel estimates.

Lemma 3: Consider an uplink system with a set of normalized uplink combining vectors and uplink power coefficients p_k, for $k = 1, \ldots, K$. Let $\{\widetilde{\mathbf{a}}_k : k = 1, \ldots, K\}$ be the unit-norm LSFD weighting vectors. If the LSFP weighting vectors in downlink are selected as

$$\widetilde{\mathbf{b}}_k = \sqrt{\rho_k} \widetilde{\mathbf{a}}_k, \tag{41}$$

and the local precoding vectors are selected identically to the normalized uplink combining vectors as in (35), then each UE can achieve the same downlink SINR as its uplink SINR $\widetilde{\mathsf{SINR}}_k^{\mathrm{ul}}$. More precisely,

$$\mathsf{SINR}_k^{\mathrm{dl}} = \widetilde{\mathsf{SINR}}_k^{\mathrm{ul}} = \frac{|\widetilde{\mathbf{a}}_k^{\mathrm{H}} \boldsymbol{\xi}_k|^2}{\widetilde{\mathbf{a}}_k^{\mathrm{H}} (\boldsymbol{\Delta}_k - \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{\mathrm{H}}) \widetilde{\mathbf{a}}_k}, \quad k = 1, \ldots, K, \tag{42}$$

for a certain power allocation policy $\{\rho_k : k = 1, \ldots, K\}$ that satisfies $\sum_{k=1}^{K} \rho_k \leq \sum_{k=1}^{K} p_k$, where $\boldsymbol{\Delta}_k$ and $\boldsymbol{\xi}_k$ are given as in (11)-(12).

Proof: The proof follows the same approach as in [32], but for the long-term LSFP and LSFD vectors. The details are relegated to Appendix B for completeness. ∎

Lemma 3 guarantees that equal effective SINRs can be achieved in the uplink and downlink, if the power allocation coefficients are selected in a unique manner, and the LSFD and LSFP vectors are identical. This implies that if we optimize the LSFD weights properly, which we have already studied how to do, we can use the same solution for LSFP. In particular, the sparse LSFD design turns into a sparse LSFP design that provides joint AP-UE assignment and downlink power allocation. There is only one caveat: the downlink power allocation suggested by the duality result might not comply with the per-AP transmit power constraints. This can be settled by appropriate centralized downlink power allocation schemes (i.e., selecting the proper per-AP power coefficients $\{\rho_k\}$), elaborated later in this section. Moreover, since the LSFD and LSFP vectors are computed at the CPU based on the long-term channel statistics and regarded as quasi-static for many time-frequency coherence blocks, the practical fronthaul links would be able to support our proposed distributed processing schemes.

Note that the uplink effective SINR in (42) is a generalized Rayleigh quotient with respect to $\widetilde{\mathbf{a}}_k$ and, thus, allows computing the LSFD vector $\widetilde{\mathbf{a}}_k^{\mathrm{opt}}$ that maximizes $\widetilde{\mathsf{SINR}}_k^{\mathrm{ul}}$ as in (13), i.e.,

$$\widetilde{\mathbf{a}}_k^{\mathrm{opt}} = \widetilde{c}_k \boldsymbol{\Delta}_k^{-1} \boldsymbol{\xi}_k \tag{43}$$

where $\widetilde{c}_k \in \mathbb{C}$ being an arbitrary non-zero scaling coefficient. Then according to (41), we have

$$\widetilde{\mathbf{b}}_k = \sqrt{\rho_k} \frac{\widetilde{\mathbf{a}}_k^{\mathrm{opt}}}{\|\widetilde{\mathbf{a}}_k^{\mathrm{opt}}\|_2}, \quad k = 1, \ldots, K, \tag{44}$$

which are referred as the *V-LSFP* vectors since we use the optimized LSFD vectors but apply a good but heuristic power allocation.

Consequentially, the virtual uplink MSE of UE k becomes

$$\widetilde{\mathsf{MSE}}_k^{\mathrm{ul}} = \widetilde{\mathbf{a}}_k^{\mathrm{H}} \boldsymbol{\Delta}_k \widetilde{\mathbf{a}}_k - 2\sqrt{p_k} \Re\left(\widetilde{\mathbf{a}}_k^{\mathrm{H}} \boldsymbol{\xi}_k\right) + p_k \tag{45}$$

which is minimized by the virtual LSFD vector in (43) with $\widetilde{c}_k = \sqrt{p_k}$ and, thus, implies that the virtual LSFD vector $\widetilde{\mathbf{a}}_k^{\mathrm{opt}}$ minimizes $\widetilde{\mathsf{MSE}}_k^{\mathrm{ul}}$ as

$$\widetilde{\mathbf{a}}_k^{\mathrm{opt}} = \underset{\widetilde{\mathbf{a}}_k \in \mathbb{C}^L}{\arg\min}\ \widetilde{\mathsf{MSE}}_k^{\mathrm{ul}}. \tag{46}$$

Similar to the uplink MSE in (14), the virtual uplink MSEs also only depend on the UE's own virtual LSFD vector $\widetilde{\mathbf{a}}_k$, which means one can find the optimal collective virtual LSFD vector

$$\widetilde{\mathbf{a}}^{\mathrm{opt}} = \underset{\widetilde{\mathbf{a}} \in \mathbb{C}^{KL}}{\arg\min} \sum_{k=1}^{K} \widetilde{\mathsf{MSE}}_k^{\mathrm{ul}} \tag{47}$$

that minimizes the virtual uplink sum MSE of all UEs as $\widetilde{\mathbf{a}}^{\mathrm{opt}} = [(\widetilde{\mathbf{a}}_1^{\mathrm{opt}})^{\mathrm{T}}, \ldots, (\widetilde{\mathbf{a}}_K^{\mathrm{opt}})^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{KL}$ where $\{\widetilde{\mathbf{a}}_k^{\mathrm{opt}} : k = 1, \ldots, K\}$ are obtained by simultaneously solving (46).

A. Centralized Downlink Power Allocation

Recall from (37) that ω_{kl} indicates the fraction of ρ_k that will be sent from AP l. Hence, the power constraint at AP l is be formulated as

$$\sum_{k\in\mathcal{D}_l} \rho_k \left|\omega_{kl}\right|^2 \leq \rho_{\max}, \tag{48}$$

where $\rho_{\max}$ is the maximal transmit power of an AP. Given $\{\omega_{kl}\}$ are already determined, the algorithms for centralized downlink power allocation can be found in [14]. One good scalable option satisfying the per-AP transmit power constraints, where the computational complexity does not grow with the number of UEs, is given as [14]

$$\rho_k = \rho_{\max} \frac{\left(\sum_{l\in\mathcal{M}_k} \beta_{kl}^{\vartheta}\right)^{\kappa} \varpi_k^{-\mu}}{\max_{j\in\mathcal{M}_k} \sum_{i\in\mathcal{D}_j} \left(\sum_{l\in\mathcal{M}_i} \beta_{il}^{\vartheta}\right)^{\kappa} \varpi_i^{1-\mu}}, \tag{49}$$

where we reshape β_{kl} with the exponent ϑ and $\varpi_i = \max_{l\in\mathcal{M}_i} |\omega_{il}|^2$ is the largest fraction of ρ_i that any of the serving APs can be assigned to transmit (see (37)), exponent $\kappa \in [-1, 1]$ determines the downlink power allocation behavior, and exponent $\mu \in [0, 1]$ is an additional parameter that reshapes the ratio of power allocation between different UEs. The rationale behind (49) is that $\rho_k \propto \left(\sum_{l\in\mathcal{M}_k} \beta_{kl}^{\vartheta}\right)^{\kappa} \varpi_k^{-\mu}$, which implies each serving AP of UE k should manage its power constraint as if it transmits with power $\rho_k \varpi_k^{\mu}$.

B. Sparse Optimization for the Downlink

Similar to the uplink, all values of LSFP vectors obtained by Lemma 3 are non-zero. With the observation that $\widetilde{\mathsf{MSE}}_k^{\mathrm{ul}}$ in (45) also possesses the quadratic structure in terms of the virtual LSFD vector $\widetilde{\mathbf{a}}_k$, the sparse algorithms developed in Section IV can also applied to enforce sparsity on the LSFP vectors in the downlink.

VI. Power Consumption Model and Energy Efficiency

The benefit of the proposed sparsity approach to compute the AP-UE association is that we can achieve almost the same SEs as when all APs serve all UEs, but with vastly less fronthaul signaling and signal processing complexity. In this section, we will define a generic power consumption model that can quantify these benefits. The model captures the following main components: a) the radio site power consumption including the power consumed at the UEs $\{P_k^{\mathrm{ue}} : \forall k\}$, the active APs $\{P_l^{\mathrm{ap}} : \forall l\}$, and fronthaul connections $\{P_l^{\mathrm{fh}} : \forall l\}$; and b) the CPU power consumption P_{cpu}. The total power consumption is modeled as

$$P_{\mathrm{tot}} = \sum_{k=1}^{K} P_k^{\mathrm{ue}} + \sum_{l=1}^{L} P_l^{\mathrm{ap}} + \sum_{l=1}^{L} P_l^{\mathrm{fh}} + P_{\mathrm{cpu}}. \tag{50}$$

We will now model each of these terms in detail.

The power consumption at a generic UE k is

$$P_k^{\mathrm{ue}} = P_k^{\mathrm{c,ue}} + \frac{\tau_{\mathrm{p}} p_{\mathrm{p}} + \tau_{\mathrm{u}} p_k}{\tau_{\mathrm{c}} \eta_{\mathrm{ue}}} \tag{51}$$

where $P_k^{\mathrm{c,ue}}$ is the internal circuit power and the second term includes the power consumption of uplink transmission, where p_{p} is the uplink pilot transmit power, p_k is the uplink data transmit power of UE k, and $0 < \eta_{\mathrm{ue}} \leq 1$ is the power amplifier efficiency at the UEs. $\tau_{\mathrm{p}}/\tau_{\mathrm{c}}$ and $\tau_{\mathrm{u}}/\tau_{\mathrm{c}}$ denotes the fractions of uplink pilot and uplink data transmission, respectively.

The power consumption related to AP l is

$$P_l^{\mathrm{ap}} = N P_l^{\mathrm{c,ap}} + N|\mathcal{D}_l| \cdot P_l^{\mathrm{pro}} + \frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{c}} \eta_{\mathrm{ap}}} \sum_{k\in\mathcal{D}_l} \rho_{kl}, \tag{52}$$

where $P_l^{\mathrm{c,ap}}$ is the internal circuit power per AP antenna, P_l^{pro} is the consumed power for processing the received/transmitted signal of each UE in $\mathcal{D}_l$, ρ_{kl} is the downlink data transmit power that AP l allocates to UE k, and $0 < \eta_{\mathrm{ap}} \leq 1$ is the power amplifier efficiency at the APs.

The fronthaul connections are used to transfer signals between the APs and the CPU. The power consumption of each fronthaul link is

$$P_l^{\mathrm{fh}} = P_l^{\mathrm{fix}} + \frac{\tau_{\mathrm{u}} + \tau_{\mathrm{d}}}{\tau_{\mathrm{c}}} |\mathcal{D}_l| \cdot P_l^{\mathrm{sig}}, \tag{53}$$

where P_l^{fix} is the fixed power consumption and remaining part describes the load-dependent uplink and downlink signaling, where P_l^{sig} is the signaling power per UE.

The CPU is responsible for processing the signals of all UEs, with power consumption

$$P_{\mathrm{cpu}} = P_{\mathrm{cpu}}^{\mathrm{fix}} + B \sum_{k=1}^{K} \left(\mathsf{SE}_k^{\mathrm{ul}} \cdot P_{\mathrm{cpu}}^{\mathrm{dec}} + \mathsf{SE}_k^{\mathrm{dl}} \cdot P_{\mathrm{cpu}}^{\mathrm{cod}} \right) \tag{54}$$

where $P_{\mathrm{cpu}}^{\mathrm{fix}}$ is the fixed power consumption, B is the system bandwidth, $P_{\mathrm{cpu}}^{\mathrm{dec}}$ is the energy consumption per bit for the final decoding at the CPU, and $P_{\mathrm{cpu}}^{\mathrm{cod}}$ is the energy consumption per bit for the initial encoding at the CPU. Typical values for these parameters are given in Table I.

With the defined power consumption model, the total EE (in bit/Joule) considering both uplink and downlink is given as [4] and [33]

$$\mathsf{EE} = B \cdot \sum_{k=1}^{K} \left(\mathsf{SE}_k^{\mathrm{ul}} + \mathsf{SE}_k^{\mathrm{dl}} \right) / P_{\mathrm{tot}}. \tag{55}$$

VII. Numerical Results

In this section, we quantify the performance achieved by our proposed LSF processing schemes in Section III and Section V, considering different combining and precoding schemes and AP deployment setups. Specifically, the L-MMSE and MR combiners are used for the uplink, and the L-MMSE and MR precoders are used for the downlink. We will measure performance in terms of SE, EE, and number of serving APs per UE (marked as "no. AP/UE" in the figures).

We consider two different AP deployments: a) $L = 40$ APs with $N = 4$ antennas and b) $L = 160$ APs with $N = 1$ antenna. The total number of antennas is $LN = 160$ in both cases. All APs and $K = 20$ UEs are distributed in the coverage area of 0.5×0.5 km^2 at random following an independent and uniform distribution. We use the wrap-around

TABLE I
SYSTEM PARAMETERS

Parameters	Values	Parameters	Values	Parameters	Values	Parameters	Values
B, τ_{c}, τ_{p}	20 MHz, 200, 10	η_{ue}, η_{ap}	0.4, 0.4	θ, ν	0.5, 0.5	ϑ, κ, μ	0.2, -0.4, 0.5
p_{p}, $p_{\max}$	0.1 W, 0.1 W	$\rho_{\max}$	1 W	$P_{\mathrm{cpu}}^{\mathrm{fix}}$, P_l^{fix}	5 W, 0.825 W	$P_k^{\mathrm{c,ue}}$, $P_l^{\mathrm{c,ap}}$	0.1 W, 0.2 W
P_l^{sig}	0.01 W	P_l^{pro}	0.8 W	$P_{\mathrm{cpu}}^{\mathrm{dec}}$	0.8 W/(Gbit/s)	$P_{\mathrm{cpu}}^{\mathrm{cod}}$	0.1 W/(Gbit/s)

TABLE II
THE SCHEMES AND BENCHMARKS FOR THE UPLINK

Schemes	AP-UE association	LSFD: $\mathbf{a}_k = [a_{k1}, \ldots, a_{kL}]^{\mathrm{T}}$, $k = 1, \ldots, K$
O-LSFD [15]	All APs serve all UEs.	$\mathbf{a}_k$ is optimized by (13).
P-LSFD [14]	Heuristic scheme [21]	$\mathbf{a}_k = c_k \left(\sum_{i \in \mathcal{P}_k} p_i \mathbb{E}\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\} + \sigma^2 \mathbf{I}_L \right)^{-1} \boldsymbol{\xi}_k$ where $\mathcal{P}_k = \{i : \mathcal{M}_i \cap \mathcal{M}_k \neq \emptyset,\ i = 1, \ldots, K\}$.
S-LSFD	Sparse optimization in (17) is utilized to enforce sparsity on $\mathbf{a}_k$ achieved from scheme O-LSFD.	

topology to approximate an infinitely large network. The 3GPP Urban Microcell model [34] is used to compute the large-scale propagation conditions, such as pathloss and shadow fading. The spatial correlation matrices are generated by using the Gaussian local scattering model with the azimuth and elevation angular standard deviation of 10° and 10°, respectively, as described in [14, Sec. 2.5.3]. The SE results with L-MMSE combining/precoding are obtained from Monte Carlo simulations, while the results with MR combining/precoding are analytically computed according to the closed-form expressions in [21, Cor. 2]. After obtaining the SE and AP-UE association results, the EE values were computed using (55) with our proposed power consumption model. Moreover, the convergence of our proposed proximal algorithms in Section IV are validated by comparing them to CVX SDPT3 (Ver. 2.2) [28]. We use $\tau_{\mathrm{d}} = 0$ and $\tau_{\mathrm{u}} = 0$ when evaluating the performance for the uplink and the downlink, respectively. Unless otherwise specified, all other system parameters are given in Table I and originate from [16], [33], and [35] (and reference therein).

A. Considered Schemes and Benchmarks

In the uplink, the transmit powers $\{p_k : \forall k\}$ are selected according to the fractional power control policy [14], [22]

$$p_k = p_{\max} \frac{\min_{i \in \{1,\ldots,K\}} \left(\sum_{l \in \mathcal{M}_i} \beta_{il} \right)^{\theta}}{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\theta}}, \tag{56}$$

where $p_{\max}$ is the maximal transmit power of a UE and $\theta \in [0, 1]$ determines the control behavior. $\theta = 0$ leads to equal power control and $\theta \to 1$ promotes more user fairness.

To demonstrate the performance improvements of our joint AP-UE association and LSFD, we compare the proposed S-LSFD with two benchmarks: O-LSFD and partial LSFD (P-LSFD). The details of these benchmarks are summarized in Table II.

For the downlink, the precoding vectors are computed using (35). The transmit powers can be selected in a distributed manner as (49) [21] and [29]

$$\rho_{kl} = \rho_{\max} \frac{(\beta_{kl})^{\nu}}{\sum_{i \in \mathcal{D}_l} (\beta_{il})^{\nu}} \tag{57}$$

if $k \in \mathcal{D}_l$ and otherwise $\rho_{kl} = 0$, with $\nu \in [0, 1]$ determining the power allocation behavior. $\nu = 0$ leads to equal power allocation and $\nu \to 1$ allocates more power to the UEs with better channel conditions. If the directions of the LSFP vectors $\{\mathbf{b}_k\}$ are already determined, the per-AP power coefficients can be selected in centralized manner as (49) [14].

To highlight the performance improvements of our V-LSFP using uplink-downlink duality in Lemma 3 and joint AP-UE association and LSFP, we propose several schemes, namely heuristic FPA (H-FPA), V-LSFP, partial LSFP (P-LSFP), S-LSFP, and sparse V-LSFP (SV-LSFP). We consider a benchmark where $\{\rho_{kl}, \forall k, l\}$ for $\{\mathbf{b}_k, \forall k\}$ are selected according to (57), which is referred to as scheme FPA in the numerical results. These schemes are elaborated in Table III.

B. Analysis for the Uplink

In Fig. 2, we evaluate the considered performance metrics achieved by L-MMSE combining with the multi-antenna AP setup (i.e., $L = 40, N = 4$), where the average SE, EE, and number of serving APs per UE are demonstrated in Fig. 2(a), Fig. 2(b), and Fig. 2(c), respectively. We compare the proposed S-LSFD with the benchmarks O-LSFD and P-LSFD for various values of the regularization parameters λ and γ, where $\gamma = 0$ stands for the case of EW-sparsity. The vertical scale intervals are set to emphasize how small/large the gaps are between the curves. The first observation is that the average SE decreases as λ and γ increase since the average number of serving APs per UE decreases. It is clear that although our proposed S-LSFD *slightly* reduces the SE by around 1% (for large values of λ and γ that each UE is served by its most essential APs), it significantly increases the EE. There

TABLE III
THE SCHEMES AND BENCHMARKS FOR THE DOWNLINK

Schemes	**AP-UE association**	**LSFP**: $\mathbf{b}_k = [\sqrt{\rho_{k1}}, \ldots, \sqrt{\rho_{kL}}]^{\mathrm{T}}$, $k = 1, \ldots, K$
FPA [29]	Heuristic scheme [21]	$\mathbf{b}_k = [\sqrt{\rho_{k1}}, \ldots, \sqrt{\rho_{kL}}]^{\mathrm{T}}$, where $\{\rho_{kl}\}$ are selected according to (57).
H-FPA	Heuristic scheme [21]	$\mathbf{b}_k = \sqrt{\rho_k}\frac{\boldsymbol{\rho}_k}{\|\boldsymbol{\rho}_k\|_2}$, where $\boldsymbol{\rho}_k = [\rho_{k1}, \ldots, \rho_{kL}]^{\mathrm{T}}$, $\{\rho_{kl}\}$ are selected according to (57), and $\{\rho_k\}$ are selected according to (49).
V-LSFP	All APs serve all UEs.	$\mathbf{b}_k$ is computed by (44) where $\{\rho_k\}$ are selected according to (49).
P-LSFP	Heuristic scheme [21]	$\mathbf{b}_k = \sqrt{\rho_k}\frac{\widetilde{\mathbf{a}}_k}{\|\widetilde{\mathbf{a}}_k\|_2}$, where $\widetilde{\mathbf{a}}_k = \widetilde{c}_k \left(\sum_{i \in \mathcal{P}_k} \mathbb{E}\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\} + \sigma^2 \mathbf{I}_L\right)^{-1} \boldsymbol{\zeta}_k$ and $\{\rho_k\}$ are selected according to (49).
S-LSFP	Sparse optimization in (17) is utilized to enforce sparsity on $\mathbf{b}_k$ achieved from scheme V-LSFP.	
SV-LSFP	Association achieved from scheme S-LSFD	$\mathbf{b}_k$ is computed by (44) where $\{\rho_k\}$ are selected according to (49).

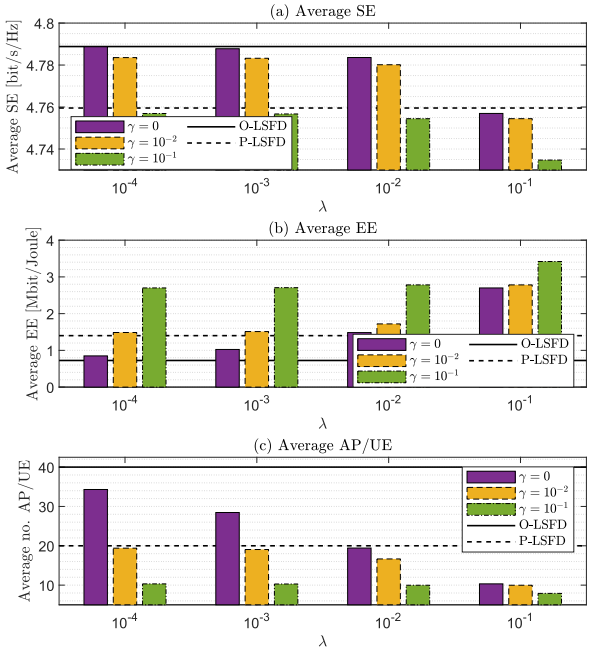


Fig. 2. Uplink average SE, EE, and number of serving APs per UE with L-MMSE combining ($L = 40, N = 4$).

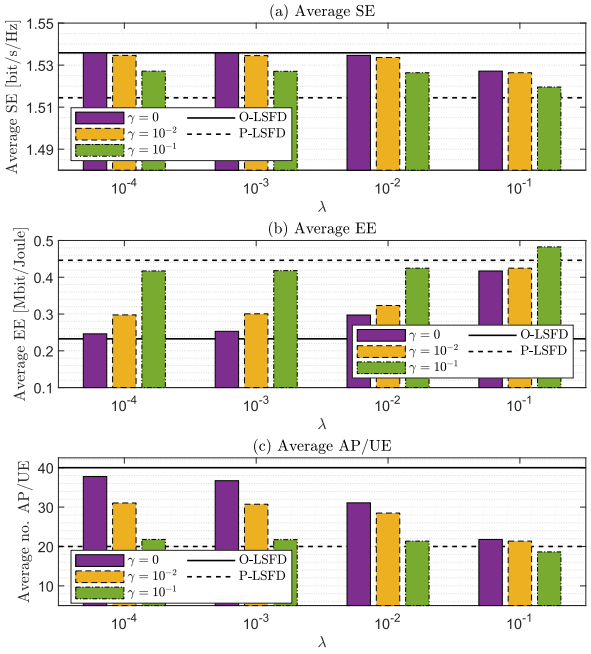


Fig. 3. Uplink average SE, EE, and number of serving APs per UE with MR combining ($L = 40, N = 4$).

is a $4\times$ EE gain compared to O-LSFD where all APs serve all UEs. Compared to P-LSFD, S-LSFD provides larger SE and similar EE by using approximately the same number of the serving APs per UE (with $\lambda = 10^{-4}, \gamma = 10^{-2}$) and also provides $1.92\times$ EE and similar SE by using half number of the serving APs per UE (with $\lambda = 10^{-1}, \gamma = 0$). The reason for this is that our joint AP-UE association and LSFD design outperform P-LSFD where the association and LSFD are performed separately. That also implies that S-LSFD is capable of making a better tradeoff between the SE and EE than P-LSFD by adjusting λ and γ.

Fig. 3 shows the results achieved by MR combining with the multi-antenna AP setup. Compared to Fig. 2, it is clear that L-MMSE combining outperforms MR regarding SE thanks to its interference suppression. Moreover, although MR may require less processing power than L-MMSE, it still cannot compensate for its disadvantage of throughput, which leads to lower EE. Similar trends in SE and EE concerning λ and γ as in Fig. 2 can be observed. It is worth noting in Fig. 2(c) and Fig. 3(c) that with the same AP deployment, MR combining benefits more from using many APs, which is reflected by having more serving APs per UE than with L-MMSE combining for all combinations of λ and γ. This is because there is so much interference when using MR that also APs that have rather weak channels to the UE can positively improve the SE (see the ranges in Fig. 2(a) and Fig. 3(a)).

Since the influence of the regularization parameter γ is similar to that of λ, which has been demonstrated in Fig. 2 and Fig. 3, the following figures with respect to sparse optimization will only consider the EW-sparsity (i.e., $\gamma = 0$). Fig. 4 is dedicated to the single-antenna AP setup (i.e., $L = 160$, $N = 1$), where L-MMSE and MR combining are both used.

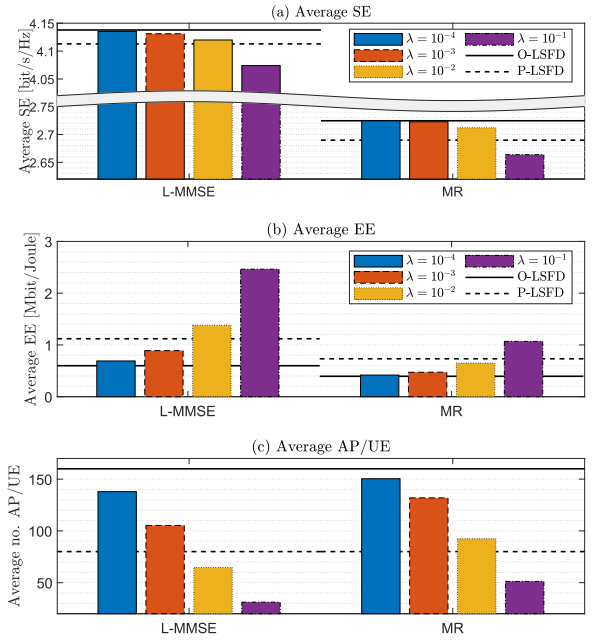


Fig. 4. Uplink average SE, EE, and number of serving APs per UE with different combiners ($\gamma = 0, L = 160, N = 1$).

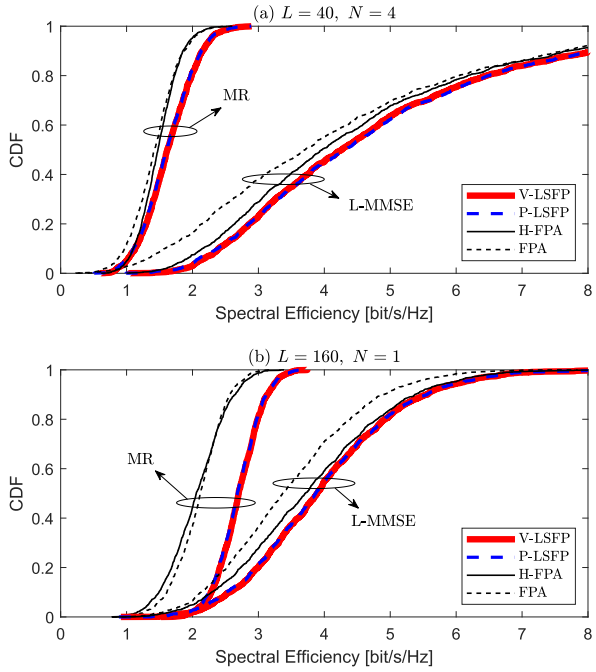


Fig. 5. Downlink SE per UE of the non-sparse schemes with different precoders and AP deployment setups.

Since the SE gaps between L-MMSE and MR is very large, we break the vertical axis in Fig. 4(a) and remove the unnecessary blank space for clear presentation. Compared to Fig. 2 and Fig. 3, we notice that the multi-antenna AP setup outperforms the single-antenna AP setup with L-MMSE combining case while it is the opposite with MR combining. The reason is that in the L-MMSE case, the interference suppression gain enabled by multiple antennas is more beneficial than the macro-diversity gain brought by having more APs. Conversely, the macro-diversity gain dominates in the MR case, which relies on it for avoiding interference. Another observation is that the EE gaps between L-MMSE and MR is larger with multi-antenna APs (between Fig. 2(b) and Fig. 3(b)) than with single-antenna APs (see Fig. 4(b)) thanks to the interference suppression.

C. Analysis for the Downlink

According to whether it involves the sparse optimization or not, the schemes considered in the downlink can be divided into two categories: the non-sparse schemes and the sparse schemes. The former includes FPA, H-FPA, V-LSFP, and P-LSFP, and the latter includes S-LSFP and SV-LSFP. We first evaluate the SE and EE performance of the non-sparse schemes to highlight the performance improvements achieved by our proposed V-LSFP design.

Fig. 5 shows the cumulative distribution function (CDF) of the downlink SE per UE. The proposed schemes V-LSFP, P-LSFP, and H-FPA are compared to the benchmark FPA with L-MMSE and MR precoding and two considered AP deployment setups. The first observation is H-FPA outperforms FPA by $1.5\times$ on 95%-likely SE thanks to the additional centralized FPA in (49). The 95%-likely SE is further improved by V-LSFP and P-LSFP to around $1.7\times$, which both exploit the uplink-downlink duality proposed in Lemma 3 to design the direction of the LSFP weighting vectors. The reason is the unit-norm virtual LSFD vectors $\{\widetilde{\mathbf{a}}_k\}$ used in V-LSFP and P-LSFP are optimized in (43) for interference suppression, and, thus, specify the fractions of ρ_k for the serving APs better than H-FPA, where the fractions of ρ_k are determined by the distributed PFA in (57). Scheme P-LSFP has a slightly lower 95%-likely SE compared to V-LSFP due to the reduced number of serving APs per UE. By comparing Fig. 5(a) and Fig. 5(b), we notice that the SE gap between the proposed schemes and the benchmark FPA is large with $L = 40, N = 4$ and shrinks with $L = 160, N = 1$ in the L-MMSE case, while it is the opposite in the MR case. This is because the L-MMSE precoder benefits from the interference suppression gain enabled by multiple antennas more than the macro-diversity gain brought by having more APs, and the MR precoder is the opposite.

The average EE of the non-sparse schemes is shown in Fig. 6 (with two precoders and two AP deployment setups), from which we observe that FPA outperforms V-LSFP where all APs serve all UEs. P-LSFP and H-FPA achieve higher EE than FPA by allocating the downlink transmit power more appropriately. When comparing the EE gaps between the two AP deployment setups, we have a similar observation of the SE gaps in Fig. 5 for the similar reason.

From Fig. 5 and Fig. 6 it is clear that V-LSFP and P-LSFP outperform the other two non-sparse schemes on SE and act as the lower and upper bound of the average EE, respectively. Therefore, to highlight the performance of the sparse LSFP schemes, we only include V-LSFP and P-LSFP into the following comparisons for concise presentation. Moreover, as we already observed, the L-MMSE precoder outperforms the MR precoder and benefits more from the multi-antenna AP setup. Thus, only the case with L-MMSE precoding and $L = 40, N = 4$ is presented.

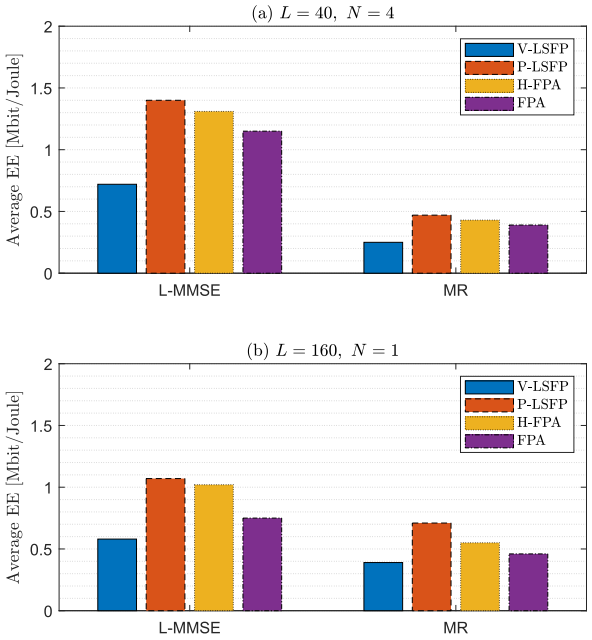


Fig. 6. Downlink average EE of the non-sparse schemes with different precoders and AP deployment setups.

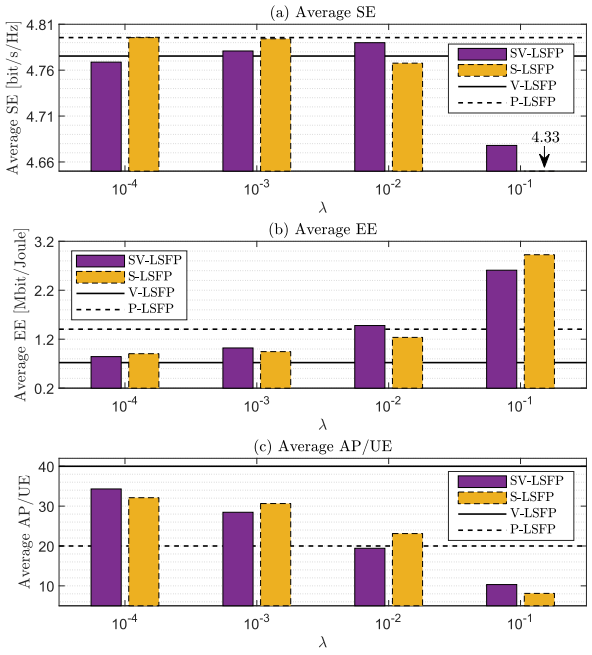


Fig. 7. Downlink average SE, EE, and number of serving APs per UE with L-MMSE precoding ($\gamma = 0, L = 40, N = 4$).

In Fig. 7, we evaluate the average SE and EE of our LSFP schemes by considering L-MMSE precoding with $L = 40, N = 4$. Unlike the uplink case in Fig. 2, V-LSFP has a lower average SE compared to its partial version P-LSFP. One reason for this result is that an appropriate transmit power allocation influenced by the AP-UE association is essential for downlink operation, where the signals from a remote AP might not contribute to the desired signal of the intended UEs, and even cause interference for the other UEs if the transmit power is not well allocated. Another reason comes from the suboptimality of the L-MMSE precoding unlike its uplink

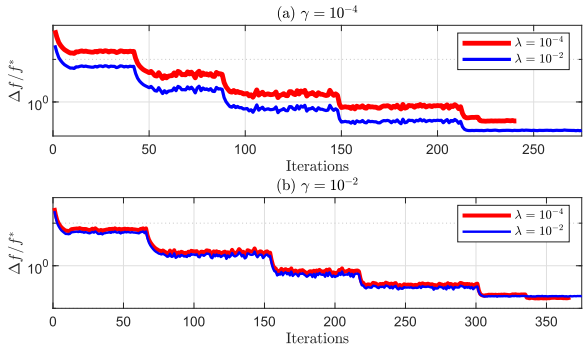


Fig. 8. Convergence accuracy with different sparsity parameters ($L = 40, N = 4$).

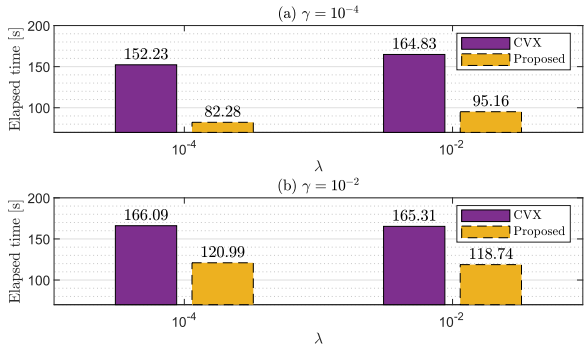


Fig. 9. Elapsed time for convergence in Fig. 8 with CVX and our proposed algorithm in Algorithm 3 ($L = 40, N = 4$).

counterpart. For the sparse schemes S-LSFP and SV-LSFP, we observe that although the sparse optimization of S-LSFP is directly performed on the downlink V-LSFP vectors, it could not maintain an absolute advantage over SV-LSFP, which exploits the sparse association from uplink for the downlink operation, on both SE and EE. In fact, these two sparse LSFP schemes are comparable with each other. The one with more serving APs per UE might win on SE but lose on EE. This is because V-LSFP is a heuristic scheme where the V-LSFP weighting vectors and the precoding vectors are computed by using uplink-downlink duality. As a consequence, the improvement of directly performing sparse optimization in the downlink is not guaranteed. When compared to the non-sparse schemes, S-LSFP and SV-LSFP are competitive with comparable SE with P-LSFP, provide higher EE than V-LSFP, and a better tradeoff between the SE and EE. In addition, we notice that the average SE in Fig. 7(a) is unimodal with respect to λ, which implies that there exists a value of λ that provides maximum average SE.

D. Algorithmic Convergence

We consider two metrics to validate the convergence of our proposed proximal algorithm with randomly generated matrices and vectors in the optimization problems: the accuracy and the elapsed time. The accuracy is shown in Fig. 8 and is defined as $\Delta f/f^*$, which is the function value difference $\Delta f = f - f^*$ normalized by the "optimal" value f^* obtained by CVX. The elapsed times for convergence with different sparsity parameters are given in Fig. 9, where

the CVX solver is considered as the benchmark for our Algorithm 3. Fig. 8 validates the correctness of our algorithm by showing the accuracy of 10^{-4}. Also, we observe that the proximal algorithm converges faster with a larger λ, where the staircase comes from the warm-restart operation. The results in Fig. 9 demonstrate the effectiveness of our algorithm where the elapsed time of our algorithm is much less than that of CVX, especially when λ and γ are small. And this advantage in terms of effectiveness will grow in large-scale networks.

VIII. Conclusion

This paper developed a joint optimization framework for the AP-UE association and distributed decoding/precoding in CF mMIMO systems. It is based on formulating and solving two sparsity-inducing MSE-minimizing problems that aim for EW and GW sparsity, respectively. The former limits the number of UEs served on average by each AP and the latter also encourages APs to not serve any UEs when not essential, both in an effort to reduce signaling and computations to improve the EE. We developed proximal algorithms to solve the formulated sparsity problems given the predetermined sparsity parameters, where the BCD approach was used for the GW case. Based on the sparse optimization, we proposed the S-LSFD scheme for the uplink. For the downlink, we first proposed the new V-LSFP by using uplink-downlink duality, which achieves a good heuristics distributed precoding. By only considering the UEs with common serving APs during the interference suppression of V-LSFP, we proposed the P-LSFP where each UE is served by a limited number of APs instead all of them. Then, we proposed the S-LSFP where the sparse association is directly obtained in the downlink, and the SV-LSFP where the association is obtained by S-LSFD in the uplink and then used as a priori for P-LSFP in the downlink.

The numerical results demonstrated that our joint optimization of AP-UE association and signal processing outperforms the existing approach, in which these operations are performed separately. The gain is especially large when using L-MMSE combining with multi-antenna APs. For example, in the uplink, the proposed S-LSFD achieved $4\times$ higher EE than O-LSFD, while only losing 1% in SE. S-LSFD achieved a $1.92\times$ EE gain and similar SE by using half number of serving APs per UE. For the downlink, our H-FPA achieved $1.5\times$ 95%-likely SE compared to FPA by using further power allocation (with L-MMSE precoder and multi-antenna APs). Under the same setup, our V-LSFP and P-LSFP increased this 95%-likely SE advantage to $1.7\times$ thanks to the virtual uplink optimization. When considering EE, FPA outperforms V-LSFP while falling behind P-LSFP and H-FPA, where the former shows higher EE. The sparse optimization also works well in the downlink where S-LSFP and SV-LSFP achieved comparable SE with P-LSFP, higher EE than V-LSFP, and a better tradeoff between the SE and EE. The comparison between S-LSFP and SV-LSFP implies that the sparse associations in the uplink and downlink are analogical when the proposed uplink-downlink duality is used.

Appendix A
Proof of Lemma 2

Since (30) is convex, the optimal solution $\boldsymbol{x}_l^*$ is characterized by the subgradient equation

$$G(\boldsymbol{x}_l^n) - \boldsymbol{x}_l^* = \mu\gamma\partial\|\boldsymbol{x}_l^*\|_2 + \mu\lambda\partial\|\boldsymbol{x}_l^*\|_1, \tag{58}$$

where

$$\partial\|\boldsymbol{x}_l^*\|_2 = \begin{cases} \dfrac{\boldsymbol{x}_l^*}{\|\boldsymbol{x}_l^*\|_2}, & \text{if } \boldsymbol{x}_l^* \neq \mathbf{0} \\ \in \{\mathbf{u} : \|\mathbf{u}\|_2 \leq 1\}, & \text{otherwise} \end{cases} \tag{59}$$

and

$$[\partial\|\boldsymbol{x}_l^*\|_1]_i = \begin{cases} \operatorname{sign}([\boldsymbol{x}_l^*]_i), & \text{if } [\boldsymbol{x}_l^*]_i \neq 0 \\ \in \{u : |u| \leq 1\}, & \text{otherwise} \end{cases}, \quad i = 1, \ldots, 2K, \tag{60}$$

are the subgradients of $\|\boldsymbol{x}_l^*\|_2$ and $\|\boldsymbol{x}_l^*\|_1$, respectively. After some algebraic manipulations, we notice that the subgradient equations are satisfied with $\boldsymbol{x}_l^* = \mathbf{0}$ if $\|\operatorname{Prox}_{\mu\lambda,\ell_1}(G(\boldsymbol{x}_l^n))\|_2 \leq \mu\gamma$, and otherwise $\boldsymbol{x}_l^*$ satisfy $\|\operatorname{Prox}_{\mu\lambda,\ell_1}(G(\boldsymbol{x}_l^n))\|_2 = \|\boldsymbol{x}_l^*\|_2 + \mu\gamma$. Then with the definition of the proximal operator of the ℓ_2-norm in (32), we obtain the closed-form expression of $\boldsymbol{x}_l^*$ as in (31) and this concludes the proof of Lemma 2.

Appendix B
Proof of Lemma 3

We prove our claim by first noting that the uplink CF SINR given in (10) has the same form as in the uplink SINR in [32, Eq. (10)] when the LSFD vectors take the role of receive beamforming weight vectors. Similarly the downlink CF SINR given in (40) has the same form as in the downlink SINR in [32, below Eq. (14)] when the normalized LSFP vectors $\mathbf{b}_k/\sqrt{\rho_k}$ take the role of transmit beamforming weight vectors. Now, consider the virtual uplink system with the SINRs $\widetilde{\mathsf{SINR}}_k^{\mathrm{ul}}$ in (42). Then, the problem of minimizing total downlink power $\sum_{k=1}^K \rho_k$ under the downlink SINR constraints $\mathsf{SINR}_k^{\mathrm{dl}} \geq \widetilde{\mathsf{SINR}}_k^{\mathrm{ul}}$ is feasible and at the optimal solution, $\mathsf{SINR}_k^{\mathrm{dl}} = \widetilde{\mathsf{SINR}}_k^{\mathrm{ul}}$ is achievable in the downlink when the LSFP vectors are selected as in (41). The optimal objective value is $\sum_{k=1}^K \rho_k \leq \sum_{k=1}^K p_k$. The equality is achieved when the power coefficients p_k and the LSFD vectors $\widetilde{\mathbf{a}}_k$ are the optimal solutions to the uplink power minimization problem, as proved in detail in [32, p. 1442].

References

[1] S. Chen, J. Zhang, E. Björnson, Ö. T. Demir, and B. Ai, "Sparse large-scale fading decoding in cell-free massive MIMO systems," in *Proc. IEEE 23rd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.

[2] C. V. N. Index, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022 White Paper*. San Jose, CA, USA: Cisco, 2019.

[3] S. Han and S. Bian, "Energy-efficient 5G for a greener future," *Nature Electron.*, vol. 3, no. 4, pp. 182–184, Apr. 2020.

[4] 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.

[5] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[6] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[7] J. G. Andrews et al., "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[8] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[9] S. Buzzi, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 697–709, Apr. 2016.

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

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

[12] S. Chen, J. Zhang, J. Zhang, E. Björnson, and B. Ai, "A survey on user-centric cell-free massive MIMO systems," *Digit. Commun. Netw.*, vol. 8, no. 5, pp. 695–719, Oct. 2022.

[13] Ö. T. Demir, M. Masoudi, E. Björnson, and C. Cavdar, "Cell-free massive MIMO in virtualized CRAN: How to minimize the total network power?" in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 159–164.

[14] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021.

[15] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 203–207.

[16] E. Björnson 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.

[17] A. Adhikary, A. Ashikhmin, and T. L. Marzetta, "Uplink interference reduction in large-scale antenna systems," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2194–2206, May 2017.

[18] A. Ashikhmin, L. Li, and T. L. Marzetta, "Interference reduction in multi-cell massive MIMO systems with large-scale fading precoding," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6340–6361, Sep. 2018.

[19] Ö. T. Demir and E. Björnson, "Large-scale fading precoding for spatially correlated Rician fading with phase shifts," 2020, *arXiv:2006.14267*.

[20] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

[21] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[22] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086–1100, Apr. 2021.

[23] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Cell-free massive MIMO with large-scale fading decoding and dynamic cooperation clustering," in *Proc. 24th Int. ITG Workshop Smart Antennas (WSA)*, Nov. 2021, pp. 1–6.

[24] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[25] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[26] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.

[27] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798–6812, Oct. 2020.

[28] CVX Research Inc. (2015). *CVX: MATLAB Software for Disciplined Convex Programming, Academic Users*. [Online]. Available: http://cvxr.com/cvx/

[29] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[30] I. Rish and G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[31] S. Noah, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Stat.*, vol. 22, no. 2, pp. 231–245, May 2013.

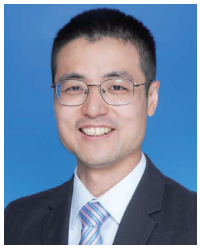
[32] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, Oct. 1998.

[33] 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.

[34] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, document TS 36.814, 3GPP, Mar. 2017.

[35] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

Shuaifei Chen (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from Beijing Jiaotong University, China, in 2018, where he is currently pursuing the Ph.D. degree. From 2019 to 2020, he was with the Department of Communication Systems, Linköping University, Sweden. From 2021 to 2022, he was with the Division of Communication Systems, KTH Royal Institute of Technology, Sweden. His research interests include signal processing and resource allocation for wireless communications, cell-free massive MIMO, and electromagnetic information theory for 6G multiple antenna technologies. He was recognized as an Exemplary Reviewer of IEEE TRANSACTIONS ON COMMUNICATIONS in 2021.



Jiayi Zhang (Senior Member, IEEE) received the Ph.D. degree in communication engineering from Beijing Jiaotong University, China, in 2014.

From 2014 to 2015, he was a Humboldt Research Fellow with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen–Nürnberg (FAU), Germany. From 2014 to 2016, he was also a Post-Doctoral Research Associate with the Department of Electronic Engineering, Tsinghua University, China. Since 2016, he has been a Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. His current research interests include cell-free massive MIMO, reconfigurable intelligent surface (RIS), communication theory, and applied mathematics. He received the Best Paper Awards from the WCSP 2017 and APCC 2017, the URSI Young Scientist Award in 2020, and the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2020. He was recognized as an Exemplary Reviewer of the IEEE COMMUNICATIONS LETTERS from 2015 to 2017 and the IEEE TRANSACTIONS ON COMMUNICATIONS from 2017 to 2019. He was the Lead Guest Editor of the Special Issue on "Multiple Antenna Technologies for Beyond 5G" of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and an Editor of IEEE COMMUNICATIONS LETTERS from 2017 to 2021. He currently serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.

Emil Björnson (Fellow, IEEE) received the M.S. degree in engineering mathematics from Lund University, Sweden, in 2007, and the Ph.D. degree in telecommunications from the KTH Royal Institute of Technology, Sweden, in 2011.

From 2012 to 2014, he was a Post-Doctoral Researcher at the Alcatel-Lucent Chair on Flexible Radio, SUPELEC, France. From 2014 to 2021, he held different professor positions at Linköping University, Sweden. He was a Visiting Full Professor at KTH from 2020 to 2021, before obtaining a tenured position in 2022. He is currently a Full Professor of wireless communication with the KTH Royal Institute of Technology. He has authored the textbooks *Optimal Resource Allocation in Coordinated Multi-Cell Systems* (2013), *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency* (2017), and *Foundations of User-Centric Cell-Free Massive MIMO* (2021). He has performed MIMO research for 16 years, his papers have received more than 23000 citations, and he has filed more than 20 patent applications. He is a host of the podcast Wireless Future and has a popular YouTube channel called Wireless Future. He is dedicated to reproducible research and has made a large amount of simulation code publicly available. His research interests include MIMO communications, radio resource allocation, machine learning for communications, and energy efficiency.

Prof. Björnson is a Wallenberg Academy Fellow, a Digital Futures Fellow, and an SSF Future Research Leader. He has received the 2014 Outstanding Young Researcher Award from IEEE ComSoc EMEA, the 2015 Ingvar Carlsson Award, the 2016 Best Ph.D. Award from EURASIP, the 2018 and 2022 IEEE Marconi Prize Paper Awards in Wireless Communications, the 2019 EURASIP Early Career Award, the 2019 IEEE Communications Society Fred W. Ellersick Prize, the 2019 IEEE Signal Processing Magazine Best Column Award, the 2020 Pierre-Simon Laplace Early Career Technical Achievement Award, the 2020 CTTC Early Achievement Award, the 2021 IEEE ComSoc RCC Early Achievement Award, and the 2023 IEEE ComSoc Outstanding Paper Award. He also coauthored papers that received Best Paper Awards at the conferences, including WCSP 2009, the IEEE CAMSAP 2011, the IEEE SAM 2014, the IEEE WCNC 2014, the IEEE ICC 2015, and the WCSP 2017. He is an Area Editor of *IEEE Signal Processing Magazine*.

Özlem Tuğfe Demir (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2012, 2014, and 2018, respectively. She was a Post-Doctoral Researcher with Linköping University, Sweden, from 2019 to 2020, and the KTH Royal Institute of Technology, Sweden, from 2021 to 2022. She is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, TOBB University of Economics and Technology, Ankara. She has authored the textbook *Foundations of User-Centric Cell-Free Massive MIMO* (2021). Her research interests include signal processing and optimization in wireless communications, massive MIMO, cell-free massive MIMO, beyond 5G multiple antenna technologies, reconfigurable intelligent surfaces, machine learning for communications, mobile data analysis, and green mobile networks.

Bo Ai (Fellow, IEEE) received the M.S. and Ph.D. degrees from Xidian University, China. He was an Excellent Post-Doctoral Research Fellow at Tsinghua University, Beijing, China, in 2007.

He was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA, in 2015. He is currently a Full Professor with Beijing Jiaotong University, where he is the Dean of the School of Electronic and Information Engineering, the Deputy Director of the State Key Laboratory of Rail Traffic Control and Safety, and the Deputy Director of the International Joint Research Center. He is one of the directors for Beijing Urban Rail Operation Control System International Science and Technology Cooperation Base, and the Backbone Member of the Innovative Engineering based jointly granted by the Chinese Ministry of Education and the State Administration of Foreign Experts Affairs. He is the research team leader of 26 national projects. He holds 26 invention patents. He has authored or coauthored eight books and authored over 300 academic research articles in his research area. Five papers have been the ESI highly cited paper. He has been notified by the Council of Canadian Academies that based on the Scopus database, he has been listed as one of the top 1% authors in his field all over the world. He has also been feature interviewed by the IET Electronics Letters. His research interests include the research and applications of channel measurement and channel modeling and dedicated mobile communications for rail traffic systems.

Dr. Ai is a fellow of The Institution of Engineering and Technology (IET) and an IEEE VTS Distinguished Lecturer. He received the Distinguished Youth Foundation and the Excellent Youth Foundation from the National Natural Science Foundation of China, the Qiushi Outstanding Youth Award by the Hong Kong Qiushi Foundation, the New Century Talents by the Chinese Ministry of Education, the Zhan Tianyou Railway Science and Technology Award by the Chinese Ministry of Railways, and the Science and Technology New Star by the Beijing Municipal Science and Technology Commission. He has won some important scientific research prizes. He is an IEEE VTS Beijing Chapter Vice Chair and an IEEE BTS Xian Chapter Chair. He was a co-chair or a session/track chair of many international conferences. He is an Associate Editor of the IEEE ANTENNAS AND WIRELESS PROPAGATION LETTERS and the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, and an Editorial Committee Member of the *Wireless Personal Communications Journal*. He is the Lead Guest Editor of Special Issues on the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the *IEEE Antennas and Propagations Letters*, and the INTERNATIONAL JOURNAL ON ANTENNAS AND PROPAGATIONS.