Unlocking the Potential of Local CSI in Cell-Free Networks with Channel Aging and Fronthaul Delays

Lorenzo Miretti and Sławomir Stańczak
Technische Universität Berlin and *Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institut*, Berlin, Germany
miretti@tu-berlin.de, slawomir.stanczak@hhi.fraunhofer.de

***Abstract*—It is generally believed that downlink cell-free networks perform best under centralized implementations where the local channel state information (CSI) acquired by the access-points (AP) is forwarded to one or more central processing units (CPU) for the computation of the joint precoders based on global CSI. However, mostly due to limited fronthaul capabilities, this procedure incurs some delay that may lead to partially outdated precoding decisions and hence performance degradation. In some scenarios, this may even lead to worse performance than distributed implementations where the precoders are locally computed by the APs based on partial yet timely local CSI. To address this issue, this study considers the problem of robust precoding design merging the benefits of timely local CSI and delayed global CSI. As main result, we provide a novel distributed precoding design based on the recently proposed *team* minimum mean-square error method. As a byproduct, we also obtain novel insights related to the AP-CPU functional split problem. Our main conclusion, corroborated by simulations, is that the opportunity of performing some local precoding computations at the APs should not be neglected, even in centralized implementations.**

I. Introduction

Cell-free massive MIMO is one of the most promising candidate technologies for enhancing the performance of future generation wireless networks [1]. Most of the related current research effort focuses on the development of practical methods and architectures for turning the known theoretical gains of coordinated multi-point concepts into commercially attractive solutions [2]–[8]. Of particular relevance is the debate on the type of joint precoding and combining implementation, and on its impact on the functional split problem in cloud radio access network (C-RAN) architectures [9].

More specifically, taken aside promising yet exotic schemes based, e.g., on sequential processing over serial fronthauls [10], [11] or iterative bidirectional over-the-air processing [12], this debate is essentially centered around the comparison between fully distributed and centralized implementations. In fully distributed implementations, the access points (AP) locally compute their precoders and combiners based on local channel state information (CSI) only [2]. In contrast, in centralized implementations, these functions are moved to one or more central processing units (CPU) endowed with global CSI [1]. Hence, due to their ability to form joint precoders and combiners based on a broader view of the channel state, centralized implementations are often considered superior, especially in terms of spectral efficiency. However, the theoretical superiority of centralized implementations is typically shown under ideal assumptions such as those related to fronthaul capabilities, which can be challenged in many practical deployments.

Against this background, this paper studies centralized downlink cell-free networks with fronthaul delays, which is a key impairment in real-world settings. In particular, we focus on scenarios where, due to fronthaul limitations and mobility, the delay incurred by the CPUs in collecting global CSI, computing the precoders, and forwarding the result is non-negligible with respect to channel aging [13]–[15]. Intuitively, in these scenarios, centralized implementations experience performance degradation and may be even outperformed by fully distributed implementations with precoders formed using partial yet more timely local CSI. To adress this issue, we formulate a distributed precoding design problem that aims to jointly exploit timely local CSI and delayed global CSI. To the best of our knowledge, this is the first time that a similar problem is addressed in the literature. Then, we derive an optimal solution based on a novel application of the recent *team* theoretical framework [11]. In addition, we discuss the structure of the optimal solution and related practical implementation aspects in C-RAN architectures. Interestingly, our theoretical and numerical results show that carefully designed implementations that delegate the computation of at least a portion of the precoders to the APs may significantly outperform both centralized and fully distributed implementations, even under pedestrian mobility and relatively small delays.

Notation: We denote by $\mathbb{R}_{++}$ the set of positive reals. The Euclidean and Frobenius norms are denoted by $\|\cdot\|$ and $\|\cdot\|_{\mathsf{F}}$, respectively. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. We denote by $\mathcal{H}^K$ the set of random vectors, i.e., K-tuples of Σ-measurable functions $\Omega \to \mathbb{C}$, satisfying $(\forall \mathbb{x} \in \mathcal{H}^K)$ $\mathsf{E}[\|\mathbb{x}\|^2] < \infty$. Given a random variable $X \in \mathcal{H}$, we denote by $\mathsf{E}[X]$ and $\mathsf{V}(X)$ its expected value and variance, respectively.

L. Miretti and S. Stańczak acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt." Joint project 6G-RIC, project identification number: 16KISK020K and 16KISK030.

II. Problem statement

A. Basic definitions and assumptions

We consider the downlink of a cell-free wireless network [1] composed of L APs indexed by $\mathcal{L} := \{1, \dots, L\}$, each of them equipped with N antennas, and K single-antenna user equipments (UEs) indexed by $\mathcal{K} := \{1, \dots, K\}$. By assuming a standard flat-fading channel model for each time-frequency resource element, we denote an arbitrary realization of the $(NL \times K)$-dimensional global channel matrix by

$$\mathbb{H} := \begin{bmatrix} \mathbb{h}_1 & \dots & \mathbb{h}_K \end{bmatrix} = \begin{bmatrix} \mathbb{H}_1 \\ \vdots \\ \mathbb{H}_L \end{bmatrix} = \begin{bmatrix} \mathbb{h}_{1,1} & \dots & \mathbb{h}_{1,K} \\ \vdots & \ddots & \vdots \\ \mathbb{h}_{L,1} & \dots & \mathbb{h}_{L,K} \end{bmatrix},$$

where $\mathbb{h}_{l,k} \in \mathcal{H}^N$ is a random vector modeling the fading state between AP $l \in \mathcal{L}$ and UE $k \in \mathcal{K}$. As customary in the channel aging literature [13], [14], we assume that the random channel realizations $\mathbb{H}$ evolve over time according to a stationary and ergodic discrete-time random process $\{\mathbb{H}[t]\}_{t\in\mathbb{Z}}$, without any further specific assumption on the time correlation. In addition, we assume that the portions $\{\mathbb{h}_{l,k}[t]\}_{t\in\mathbb{Z}}$ of $\{\mathbb{H}[t]\}_{t\in\mathbb{Z}}$ corresponding to different AP-UE pairs are mutually independent random processes.

Similarly, by focusing on simple multi-user cooperative transmission techniques based on linear precoding and on treating interference as noise [1], we denote an arbitrary realization of the $(NL \times K)$-dimensional joint precoding matrix by

$$\mathbb{T} = \begin{bmatrix} \mathbb{t}_1 & \dots & \mathbb{t}_K \end{bmatrix} = \begin{bmatrix} \mathbb{T}_1 \\ \vdots \\ \mathbb{T}_L \end{bmatrix} = \begin{bmatrix} \mathbb{t}_{1,1} & \dots & \mathbb{t}_{1,K} \\ \vdots & \ddots & \vdots \\ \mathbb{t}_{L,1} & \dots & \mathbb{t}_{L,K} \end{bmatrix},$$

where $\mathbb{t}_{l,k} \in \mathcal{H}^N$ is a linear precoding vector applied by AP $l \in \mathcal{L}$ to the coded and modulated data stream for UE $k \in \mathcal{K}$. The joint precoding matrix $\mathbb{T}$ evolves over time according to a random process $\{\mathbb{T}[t]\}_{t\in\mathbb{Z}}$, where $\mathbb{T}[t]$ is adapted to the random channel realization $\mathbb{H}[t]$ based on the available channel state information (CSI) at time $t \in \mathbb{Z}$. In particular, we focus on the case where the submatrices $\mathbb{T}_l[t]$ of $\mathbb{T}[t]$ corresponding to the precoding matrices applied by each AP $l \in \mathcal{L}$ may depend on different CSI. This aspect is treated in more details next.

B. Delayed CSI sharing

Canonical cell-free network models based on time-division duplex operations assume each AP $l \in \mathcal{L}$ to acquire local measurements of the downlink local channel $\mathbb{H}_l = [\mathbb{h}_{l,1} \ \dots \ \mathbb{h}_{l,K}]$ by means of uplink pilot signals. In centralized implementations [1], these local measurements are then typically forwarded by the APs to one or more central processors for the computation of the joint precoding matrix based on measurements of the global channel $\mathbb{H}$. However, this process inevitably incurs some delay and may lead to outdated precoding decisions in many practical scenarios, making distributed implementations based on timely local measurements [2] a competitive alternative. This observation is also corroborated by our simulations in Section IV for relatively small delays.

In this work, we study the impact of delayed CSI sharing on performance and robust precoding design by considering the following simplified model. We assume that each AP $l \in \mathcal{L}$ can form its precoding matrix based on perfect instantaneous knowledge of the local channel $\mathbb{H}_l[t]$, and perfect d-step delayed knowledge of the global channel $\mathbb{H}[t]$. More precisely, we assume that the precoders $\mathbb{T}_l[t]$ of AP $l \in \mathcal{L}$ at time $t \in \mathbb{Z}$ are constrained to be functions of the CSI

$$S_l[t] := (\mathbb{H}_l[t], Z[t]), \quad Z[t] := \mathbb{H}[t-d], \quad d \in \mathbb{N}. \tag{1}$$

As we will see, the key feature of the above model is that it allows us to design robust precoders that combine the benefits of (delayed) centralized interference management with the opportunity of performing timely local refinements.

Remark 1. *To avoid technical digressions, cumbersome notation, and to better focus on the essence of the problem, in this study we do not consider aspects such as channel estimation errors, user-centric network clustering, and the opportunity of storing and exploiting CSI history such as $\mathbb{H}[t-i]$ for $i > d$. However, these aspects can be easily incorporated in our model and results following the approach in [16], and will be covered in details in an extended version of this study.*

Remark 2. *Our model and main derivations do not explicitly consider centralized precoding computation, and assume a distributed system where all precoders are locally computed by the APs after a preliminary CSI sharing step. However, we remark that all computations involving $\mathbb{H}[t-d]$ only can also be implemented on a central processor. Hence, our model implicitly covers both the aforementioned centralized and fully distributed implementations, which correspond to extreme functions that discard either $\mathbb{H}_l[t]$ or $\mathbb{H}[t-d]$, respectively. More interestingly, our model also covers intermediate cases where the computation of the precoders is split among a central processor operating on the basis of global delayed CSI, and the APs operating on the basis of timely local CSI. Additional details are given in Section III-C.*

C. Team MMSE precoding

Following the approach in [11], which introduced a novel non-heuristic method for optimal distributed precoding design when the APs are endowed with different CSI, we consider the following parametric *Team MMSE* precoding problem:

$$\underset{\mathbb{T}\in\mathcal{T}}{\text{minimize}} \ \mathsf{E}\left[\|\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}^{\mathsf{H}}\mathbb{T} - \boldsymbol{I}_K\|_{\mathsf{F}}^2\right] + \sum_{l=1}^{L} \sigma_l \mathsf{E}\left[\|\mathbb{T}_l\|_{\mathsf{F}}^2\right], \tag{2}$$

where $\mathcal{T} \subseteq \mathcal{H}^{K\times LN}$ is a given *information* constraint [11], [16] induced by the CSI structure (1), $\boldsymbol{P} = \operatorname{diag}(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \dots, p_K) \in \mathbb{R}^K_{++}$, and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_L) \in \mathbb{R}^L_{++}$ are given parameters.[1] Note that, to improve readability of the paper, we omit the dependency on the time index t since $\{\mathbb{H}[t], S_1[t], \dots, S_L[t]\}_{t\in\mathbb{Z}}$ is a stationary random process. However, we remark that the impact of the delay d is still fully captured in (3) by means of the constraint set $\mathcal{T}$.

[1] Note that the Team MMSE precoding problem is equivalently formulated in [11], [16] as K separate problems for each of the K precoding vectors $(\mathbb{t}_k)_{k=1}^K$, coupled by the problem parameters $(\boldsymbol{p}, \boldsymbol{\sigma})$. The difference between [11] and [16] is that [11] focuses on the case $\boldsymbol{\sigma} = \mathbf{1}$.

Informally, the role of the constraint set $\mathcal{T}$ is to enforce the precoders of each AP $l \in \mathcal{L}$ to be functions of $S_l = (\mathbb{H}_l, Z)$ only, where S_l denotes a realization of (1) at some arbitrary time $t \in \mathbb{Z}$. As proposed in [11], [16], we formally define the constraint set $\mathcal{T}$ as follows: we let

$$(\forall k \in \mathcal{K})\ \mathbb{t}_k \in \mathcal{T}_k := \mathcal{H}_1^N \times \ldots \times \mathcal{H}_L^N,$$

where $\mathcal{H}_l^N \subseteq \mathcal{H}^N$ denotes the set of N-tuples of Σ_l-measurable functions $\Omega \to \mathbb{C}$ satisfying $(\forall \mathbb{x} \in \mathcal{H}_l^N)$ $\mathsf{E}[\|\mathbb{x}\|^2] < \infty$, and where $\Sigma_l \subseteq \Sigma$ is the sub-σ-algebra induced by the CSI $S_l = (\mathbb{H}_l, Z)$ available at AP $l \in \mathcal{L}$. In the team theoretical literature, Σ_l is also called the *information subfield* of AP l. Then, we let $\mathcal{T} := \mathcal{T}_1 \times \ldots \times \mathcal{T}_K$. The interested reader is referred to [17] for an introduction to the measure theoretical notions used in the above definitions. However, we stress that these notions are by no means required for understanding the key results of this study.

Remark 3. *Problem* (2) *can be motivated under multiple points of view. For instance, the solution to Problem* (2) *can be interpreted as the best distributed approximation of regularized channel inversion (recovered for $d = 0$), where the parameters $(\boldsymbol{p}, \boldsymbol{\sigma})$ can be tuned to balance UE priorities and APs power consumption. Furthermore, solving* (2) *corresponds to minimizing the individual MSE between the transmit and receive data-bearing symbols after linear processing over a dual uplink channel with UE uplink powers $\boldsymbol{p}$ and AP noise powers $\boldsymbol{\sigma}$. Finally, an information theoretical motivation is obtained by evaluating performance using the so-called* hardening *inner bound [1] on the ergodic capacity region. Specifically, by leveraging the known uplink-downlink duality principle for fading channels under a sum power constraint (see, e.g., [1]), [11] proves that, by choosing $\boldsymbol{\sigma} = \mathbf{1}$ and $\boldsymbol{p}$ such that $\sum_{k=1}^K p_l = P$, the solution to* (2) *is Pareto optimal, in the sense that it produces rate tuples on the boundary of the considered inner bound under a sum power constraint P (and unitary noise powers). Conversely, [11] shows that all boundary points under a sum power constraint P can be achieved by solutions to* (2) *for $\boldsymbol{\sigma} = \mathbf{1}$ and for some $\boldsymbol{p}$ such that $\sum_{k=1}^K p_l = P$. Furthermore, [16] proves that all boundary points under per-AP power constraints can be achieved by solutions to* (2) *for some $(\boldsymbol{p}, \boldsymbol{\sigma})$.*

The parameters $(\boldsymbol{p}, \boldsymbol{\sigma})$ of Problem (2) can be tuned to maximize some network utility function under some power and/or quality of service constraints as in, e.g, [16], [18], or set heuristically as for the many variants of the MMSE or regularized zero forcing precoding schemes [1]. Additional details on the tuning of these parameters are left for the extended version of this study, since they mostly relate to resource allocation and power control aspects that are not specific to the delayed CSI sharing model (1). The rest of this study is devoted to solving the Team MMSE precoding problem (2) under the considered delayed CSI sharing model (1), for arbitrary problem parameters $(\boldsymbol{p}, \boldsymbol{\sigma})$. Only our simulations in Section IV will focus on a specific example of $(\boldsymbol{p}, \boldsymbol{\sigma})$.

III. Problem solution

A. Optimal solution

Problem (2) is a functional (i.e., infinite dimensional) optimization problem belonging to the class of *team decision* problems [17], which are notoriously difficult, even when convex as in our case. The main difficulty lies in the information constraint $\mathcal{T}$, which prevents the direct application of standard methods and numerical routines for finite dimensional convex problems. However, [11] showed that Problem (2) can be mapped to a minor variation of the subclass of *quadratic* team problems [17], and, as a consequence, that the following necessary and sufficient optimality conditions hold.

Proposition 1. *For given $\boldsymbol{p} \in \mathbb{R}_{++}^K$ and $\boldsymbol{\sigma} \in \mathbb{R}_{++}^L$, Problem* (2) *admits a unique solution, which is also the unique $\mathbb{T} \in \mathcal{T}$ satisfying*

$$(\forall l \in \mathcal{L})\ \mathbb{T}_l = \mathbb{F}_l \left(\boldsymbol{I}_K - \sum_{j \in \mathcal{L} \setminus \{l\}} \boldsymbol{P}^{\frac{1}{2}} \mathsf{E}\left[\mathbb{H}_j^{\mathsf{H}} \mathbb{T}_j \middle| S_l \right] \right), \quad (3)$$

where $\mathbb{F}_l := \left(\mathbb{H}_l \boldsymbol{P} \mathbb{H}_l^{\mathsf{H}} + \sigma_l \boldsymbol{I}_N \right)^{-1} \mathbb{H}_l \boldsymbol{P}^{\frac{1}{2}} \in \mathcal{H}^{N \times K}$.

Proof. The proof follows readily by replacing $\boldsymbol{\sigma} = \mathbf{1}$ with an arbitrary $\boldsymbol{\sigma} \in \mathbb{R}_{++}^L$ from the proof of [11, Lemma 2]. Informally, (3) is obtained by minimizing the objective in (2) with respect to $\mathbb{T}_l$, and by fixing $\mathbb{T}_j$ for $j \neq l$. This gives a set of necessary optimality conditions, related to the game theoretical notion of Nash equilibrium. The key step of the proof shows that these conditions are also sufficient. □

Proposition 1 and its extensions covering channel estimation errors and user-centric network clustering are used in [11], [16] to derive optimal distributed precoders under local CSI models of the type $(\forall l \in \mathcal{L})\ S_l = \mathbb{H}_l$ for fully distributed implementations, or under sequential CSI sharing models of the type $(\forall l \in \mathcal{L})\ S_l = (\mathbb{H}_1, \ldots, \mathbb{H}_l)$ for partially distributed implementations exploiting the properties of serial fronthauls. In the following, we use Proposition 1 to derive the main result of this study, i.e., the solution to Problem (2) under the delayed CSI sharing model $(\forall l \in \mathcal{L})\ S_l = (\mathbb{H}_l, Z)$ in (1).

Proposition 2. *For given $\boldsymbol{p} \in \mathbb{R}_{++}^K$ and $\boldsymbol{\sigma} \in \mathbb{R}_{++}^L$, the unique solution to Problem* (2) *is given by*

$$(\forall l \in \mathcal{L})\ \mathbb{T}_l = \mathbb{F}_l \mathbb{C}_l, \quad (4)$$

where $\mathbb{F}_l := \left(\mathbb{H}_l \boldsymbol{P} \mathbb{H}_l^{\mathsf{H}} + \sigma_l \boldsymbol{I}_N \right)^{-1} \mathbb{H}_l \boldsymbol{P}^{\frac{1}{2}} \in \mathcal{H}^{N \times K}$, and $\mathbb{C}_l \in \mathcal{H}^{K \times K}$ is given by the unique solution to the linear system of equations

$$(\forall l \in \mathcal{L})\ \mathbb{C}_l + \sum_{j \in \mathcal{L} \setminus \{l\}} \mathsf{E}[\boldsymbol{P}^{\frac{1}{2}} \mathbb{H}_j^{\mathsf{H}} \mathbb{F}_j | Z] \mathbb{C}_j = \boldsymbol{I}_K. \quad (5)$$

Proof. (Sketch) The proof follows by verifying that (4) satisfies (3) via simple algebraic manipulations. □

B. Interpretation and computation of the optimal solution

Proposition 2 states the optimality of the two-stage precoding structure in (4), where each stage depends on a distinct portion of the CSI $S_l = (\mathbb{H}_l, Z)$. Specifically, the optimal precoder $\mathbb{T}_l$ at AP $l \in \mathcal{L}$ is composed by a first $N \times K$ *local* MMSE precoding stage $\mathbb{F}_l$ [1], function of the timely local

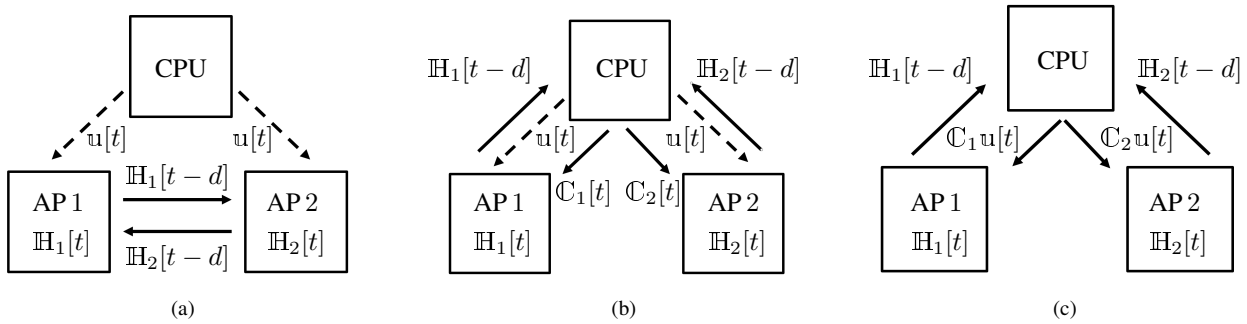

Fig. 1. Pictorial representation of possible implementations of the proposed team MMSE solution (4) in C-RAN architectures: (a) distributed precoding with CSI sharing; (b) locally refined centralized precoding (compress-before-precoding); (c) locally refined centralized precoding (compress-after-precoding).

CSI $\mathbb{H}_l$, and a second $K \times K$ precoding stage $\mathbb{C}_l$, function of the delayed global CSI Z. To better understand the dependency of $\mathbb{C}_l$ on Z, we observe that the linear system of equations (5) defining $\mathbb{C}_l$ has random coefficients $\mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l^{\mathsf{H}}\mathbb{F}_l|Z]$ which are functions of Z. In particular, each realization of the L precoding stages $(\mathbb{C}_l)_{l=1}^L$ can be obtained by solving (5) disjointly for each realization of Z. More precisely, a realization $(\boldsymbol{C}_l)_{l=1}^L$ of $(\mathbb{C}_l)_{l=1}^L$ for a given realization z of Z is given by the solution to the finite dimensional linear system of equations

$$(\forall l \in \mathcal{L})\ \boldsymbol{C}_l + \sum_{j \in \mathcal{L}\setminus\{l\}} \mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_j^{\mathsf{H}}\mathbb{F}_j|Z=z]\boldsymbol{C}_j = \boldsymbol{I}_K,$$

which can be computed using standard techniques, provided that the coefficients $\mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l^{\mathsf{H}}\mathbb{F}_l|Z=z]$ are known.

By reintroducing the time index t, we notice that these coefficients take the form $\mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l[t]^{\mathsf{H}}\mathbb{F}_l[t]|Z[t]] = \mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l[t]^{\mathsf{H}}\left(\mathbb{H}_l[t]\boldsymbol{P}\mathbb{H}_l[t]^{\mathsf{H}}+\sigma_l\boldsymbol{I}_N\right)^{-1}\mathbb{H}_l[t]\boldsymbol{P}^{\frac{1}{2}}|\mathbb{H}_l[t-d]]$, i.e., they are functions of $\mathbb{H}_l[t-d]$ defined by the conditional distribution of $\mathbb{H}_l[t]$ given $\mathbb{H}_l[t-d]$, which we recall is independent of t due to stationarity. Importantly, these coefficients can be computed in parallel for each AP. Furthermore, we notice that they can be interpreted as the optimal d-step MMSE predictors of the $K \times K$ effective channels $\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l[t]^{\mathsf{H}}\mathbb{F}_l[t]$ after local MMSE precoding. Unfortunately, closed-form expressions for the coefficients $\mathsf{E}[\boldsymbol{P}^{\frac{1}{2}}\mathbb{H}_l^{\mathsf{H}}\mathbb{F}_l|Z]$ may not be available in many practical cases. However, we remark that many approximate numerical methods taken from the vast literature on estimation/prediction theory could be potentially applied. Of particular practical interest are data driven techniques that do not require explicit knowledge of the conditional distribution of $\mathbb{H}_l[t]$ given $\mathbb{H}_l[t-d]$. For simplicity, in this work, we evaluate numerically each expectation using an empirical average over a sample set generated according to the conditional distribution of $\mathbb{H}_l[t]$ given $\mathbb{H}_l[t-d]$, assumed known. In addition, in Section III-D, we discuss some suboptimal approximations. We leave the evaluation of more advanced techniques as a promising future line of research.

C. C-RAN functional split aspects

In this section we discuss the implementation of the optimal two-stage precoding structure identified in (4) by focusing on the functional split problem in C-RAN architectures [9]. First, we recall that the considered solution in (4) is derived under the delayed CSI sharing model (1), by assuming that the precoders are computed locally at each AP after a preliminary CSI sharing step. This model can be directly mapped to the functional split depicted in Fig. 1a, where (from a physical layer perspective) the CPU only forms and forwards the K-dimensional vector of coded and modulated data streams $\mathbb{u}[t]$. This is the closest implementation to the original fully distributed cell-free massive MIMO concept proposed in [2].

However, as anticipated in Remark 2, (4) can also be implemented by splitting the computation of the two stages between the APs and the CPU, as depicted in Fig. 1b and Fig. 1c. These implementations are closer to the centralized cell-free massive MIMO concept described, e.g., in [1]. In both these functional splits, the CPU forms the precoding stages $\mathbb{C}_l[t]$ based on delayed global CSI, and the APs form their local MMSE stages $\mathbb{F}_l[t]$ based on timely local CSI. The difference between these two functional splits is that in Fig. 1b both precoding stages are applied to the data streams by the APs, while in Fig. 1c the CPU computes and forwards the K-dimensional intermediate signals $\mathbb{C}_l[t]\mathbb{u}[t]$ for each AP.[2]

Choosing the best functional split is a notoriously challenging problem encompassing many different system aspects. For instance, Fig. 1b and Fig. 1c can be respectively interpreted as novel distributed versions of the *compress-before-precoding* and *compress-after-precoding* functional splits compared in [9] in terms of fronthaul rate requirements. Interestingly, the results of this study may be used as a novel approach to extend the current literature on C-RAN functional splits covering fronthaul and processing delay requirements.

We conclude this section by pointing out that, in the current form, none of the implementations in Fig. 1 is scalable with respect to the number of UEs K. This is mostly due to the fact that the information to be shared and processed is proportional to K, as in [2]. However, we remark that this issue can be significantly mitigated by omitting the sharing and processing of information that does not contribute significantly to performance, for instance because of large path loss between a certain UE-AP pair. A popular way of implementing this idea is via the user-centric network clustering paradigm [1]. The extension of our results to this paradigm can be done as in [16], and it is left for the extended version of this study.

[2]In both cases, the delay d should be rather interpreted as the round-trip delay incurred by the two-way information sharing procedure.

D. Suboptimal solutions

1) Local precoding: By discarding the potentially useful information $\mathbb{H}[t-d]$ in (1), i.e., by letting $(\forall l \in \mathcal{L})\ S_l[t] = \mathbb{H}_l[t]$, the solution to Problem (2) is given by a variant of (4) with $(\forall l \in \mathcal{L})\ \mathbb{C}_l = \boldsymbol{C}_l$, where $(\boldsymbol{C}_l)_{l=1}^L$ are fixed (deterministic) precoding stages given by the solution to

$$(\forall l \in \mathcal{L})\ \boldsymbol{C}_l + \sum_{j \in \mathcal{L} \setminus \{l\}} \mathsf{E}[\boldsymbol{P}^{\frac{1}{2}} \mathbb{H}_j^{\mathsf{H}} \mathbb{F}_j] \boldsymbol{C}_j = \boldsymbol{I}_K. \quad (6)$$

This corresponds to the local *team MMSE* solution derived in [11], [16], which we recall is an enhanced version of the known local MMSE scheme and its variants [1].

2) Centralized (delay-tolerant) precoding: On the other extreme, by discarding $(\mathbb{H}_l[t])_{l=1}^L$ in (1), i.e., by letting $(\forall l \in \mathcal{L})\ S_l[t] = \mathbb{H}[t-d]$, the solution to Problem (2) becomes

$$\mathbb{T}[t] = \left(\hat{\mathbb{H}}[t] \boldsymbol{P} \hat{\mathbb{H}}[t]^{\mathsf{H}} + \boldsymbol{\Psi} + \boldsymbol{\Sigma} \right)^{-1} \hat{\mathbb{H}}[t] \boldsymbol{P}^{\frac{1}{2}}, \quad (7)$$

where $\hat{\mathbb{H}}[t] := \mathsf{E}[\mathbb{H}[t] | \mathbb{H}[t-d]]$, $\boldsymbol{\Psi} := \mathsf{E}[(\hat{\mathbb{H}}[t] - \mathbb{H}[t]) \boldsymbol{P} (\hat{\mathbb{H}}[t] - \mathbb{H}[t])^{\mathsf{H}}]$, and $\boldsymbol{\Sigma} := \mathrm{blkdiag}(\sigma_1 \boldsymbol{I}_N, \ldots, \sigma_L \boldsymbol{I}_N)$. This essentially corresponds to the known centralized MMSE scheme in [1] or to the delay-tolerant zero-forcing scheme based on channel prediction in [15], carefully optimized and adapted to our setup to ensure a fair comparison against (4).

3) Naïve distributed precoding: A simple baseline distributed precoding scheme that takes into account the full information in (1) is given by letting each AP $l \in \mathcal{L}$ locally compute a version of the centralized solution (7) based on its local estimate of the global channel state, obtained by replacing the submatrix $\hat{\mathbb{H}}_l[t] := \mathsf{E}[\mathbb{H}_l[t] | \mathbb{H}_l[t-d]]$ of $\hat{\mathbb{H}}[t]$ with $\mathbb{H}_l[t]$. This baseline approach was termed *naïve* precoding in the early works on distributed precoding [19], since it essentially corresponds to letting each AP believe that its information on $\mathbb{H}[t]$ is the same information at all APs.

4) Structure-aware distributed precoding: As an alternative to the above baseline distributed precoding scheme that takes into account the structure of the optimal solution, we propose a variant of (4) based on approximating the coefficients of the linear system in (5) as $\mathsf{E}[\boldsymbol{P}^{\frac{1}{2}} \mathbb{H}_l[t]^{\mathsf{H}} \mathbb{F}_l[t] | \mathbb{H}_l[t-d]] \approx$

$$\boldsymbol{P}^{\frac{1}{2}} \hat{\mathbb{H}}_l[t]^{\mathsf{H}} \left(\hat{\mathbb{H}}_l[t] \boldsymbol{P} \hat{\mathbb{H}}_l[t]^{\mathsf{H}} + \boldsymbol{\Psi}_l + \sigma_l \boldsymbol{I}_N \right)^{-1} \hat{\mathbb{H}}_l[t] \boldsymbol{P}^{\frac{1}{2}},$$

where $\boldsymbol{\Psi}_l := \mathsf{E}[(\hat{\mathbb{H}}_l[t] - \mathbb{H}_l[t]) \boldsymbol{P} (\hat{\mathbb{H}}_l[t] - \mathbb{H}_l[t])^{\mathsf{H}}]$. This is similar to the centralized approach described above, but confined to the computation of the precoding stages $(\mathbb{C}_l[t])_{l=1}^L$.

IV. Numerical simulations and conclusions

We consider a network composed by $K = 50$ UEs are uniformly distributed within a squared service area of size 0.5×0.5 km^2, and $L = 16$ regularly spaced APs with $N = 4$ antennas each. By neglecting for simplicity spatial correlation, we let $\mathbb{h}_{l,k}$ be independently distributed as $\mathbb{h}_{l,k} \sim \mathcal{CN}(\boldsymbol{0}, \gamma_{l,k} \boldsymbol{I}_N)$, where $\gamma_{l,k} > 0$ denotes the channel gain between AP l and UE k. We follow the same 3GPP-like path-loss model adopted in [1] for a 2 GHz carrier frequency:

$$\gamma_{l,k} = -36.7 \log_{10}(D_{l,k}/1\text{ m}) - 30.5 + Z_{l,k} - \sigma^2 \quad [\text{dB}],$$

where $D_{l,k}$ is the distance between AP l and UE k including a difference in height of 10 m, and $Z_{l,k} \sim \mathcal{N}(0, \rho^2)$ [dB] are shadow fading terms with deviation $\rho = 4$. The shadow fading is correlated as $\mathsf{E}[Z_{l,k} Z_{j,i}] = \rho^2 2^{-\frac{\delta_{k,i}}{9\text{ [m]}}}$ for all $l = j$ and zero otherwise, where $\delta_{k,i}$ is the distance between UE k and UE i. The noise power is $\sigma^2 = -174 + 10 \log_{10}(B/1\text{ Hz}) + F$ [dBm], where $B = 20$ MHz is the bandwidth, and $F = 7$ dB is the noise figure.

The time evolution of the channel is modeled as in [13] and many related studies by assuming that each $\{\mathbb{h}_{l,k}[t]\}_{t \in \mathbb{Z}}$ is a zero-mean stationary ergodic complex Gaussian process, where the joint distribution of $(\mathbb{h}_{l,k}[t], \mathbb{h}_{l,k}[t-d])$ is fully characterized by the autocovariance matrix $\mathsf{E}[\mathbb{h}_{l,k}[t] \mathbb{h}_{l,k}[t-d]^{\mathsf{H}}] = r_{l,k} \gamma_{l,k} \boldsymbol{I}_N$ for a given autocorrelation coefficient $r_{l,k} \in [0,1]$. The autocorrelation coefficient can be used to model the joint impact of the CSI sharing delay d and UE mobility, for example by using Clarke's model $r_{l,k} = J_0(2\pi \nu_{l,k} T d)$ as in [13], where $\nu_{l,k}$ denotes the Doppler spread for the (l,k)th AP-UE pair, and T is the symbol time. We focus on two representative scenarios obtained by letting $(\forall l \in \mathcal{L})(\forall k \in \mathcal{K})\ r_{l,k} = r \in \{0.99, 0.9\}$, which, according to Clarke's model, can be mapped to pedestrian mobility ($\nu_{l,k} \approx 10$ Hz) for all UEs and delay $Td \in \{1\text{ ms}, 10\text{ ms}\}$.

A. Figure-of-merit

We evaluate performance in terms of downlink ergodic achievable rates estimated by the hardening inner bound [1] for a given (network-wide) sum power $P = 5$ W. To facilitate the connection with the solution to Problem (2), we leverage the known uplink-downlink duality principle for fading channels (see, e.g., [1]) and compute the downlink rate of each UE $k \in \mathcal{K}$ for a given precoding scheme $\mathbb{T}$ and for some downlink power allocation policy using the equivalent expressions

$$R_k(\mathbb{T}, \boldsymbol{p}) := \log_2(1 + \mathrm{SINR}_k(\mathbb{t}_k, \boldsymbol{p})),$$

$$\mathrm{SINR}_k(\mathbb{t}_k, \boldsymbol{p}) := \frac{p_k |\mathsf{E}[\mathbb{h}_k^{\mathsf{H}} \mathbb{t}_k]|^2}{p_k \mathsf{V}(\mathbb{h}_k^{\mathsf{H}} \mathbb{t}_k) + \sum_{j \neq k} p_j \mathsf{E}[|\mathbb{h}_j^{\mathsf{H}} \mathbb{t}_k|^2] + \mathsf{E}[\|\mathbb{t}_k\|^2]},$$

corresponding to the achievable rates over a virtual uplink channel for some virtual uplink powers $\boldsymbol{p} \in \mathbb{R}_+^K$ satisfying $\sum_{k=1}^K p_k = P$. We recall that the uplink-downlink duality principle guarantees the existance of a downlink power allocation such that the same rates are also achievable in the downlink using $\mathbb{T}$. It can be shown that the above uplink rates are maximized by the solution to Problem (2) with problem parameters $\boldsymbol{p}$ equal to the uplink powers and $\boldsymbol{\sigma} = \boldsymbol{1}$ [11], [18]. As an example, we choose a fractional power allocation policy with exponent -1 [6], i.e., we let $(\forall k \in \mathcal{K})\ p_k \propto (\sum_{l=1}^L \gamma_{l,k})^{-1}$, which approximates a max-min fair policy.

B. Results

Figure 2 shows the cumulative density function of the ergodic rates achieved by different precoding schemes, approximated using 100 independent user drops and 100 independent realizations of $(\mathbb{H}[t], \mathbb{H}[t-d])$ per user drop. In particular, we compare the performance of the optimal Team MMSE solution (3) against the suboptimal solutions described in Section III-D. We remark that the schemes termed *centralized* and *local* correspond, respectively, to the best known schemes for the centralized and (fully) distributed cell-free massive MIMO

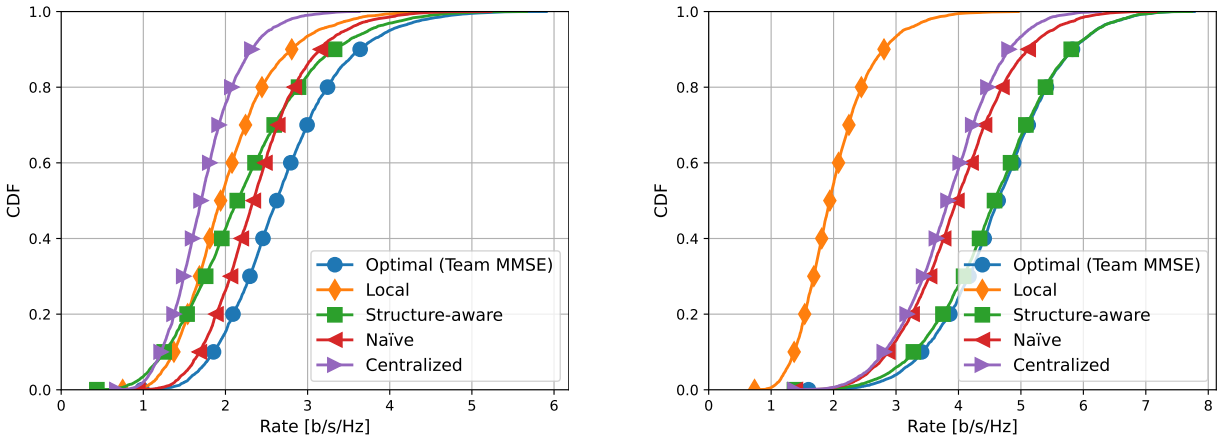

Fig. 2. Cumulative density function of downlink ergodic rates achieved by the optimal team MMSE solution (4) and the suboptimal designs in Sect. III-D, assuming pedestrian mobility and a CSI sharing delay of (a) 10 ms ($r = 0.9$) and (b) 1 ms ($r = 0.99$).

implementations described in [1]. The Team MMSE solution is computed via the numerical method described in Section III-B.

A first important observation is that, even for pedestrian mobility, a CSI sharing delay of 10 ms ($r = 0.9$, Figure 2a) can significantly degrade the performance of centralized precoding to the point where it becomes noticeably worse than the performance of local precoding. Moreover, a second important observation is that that the proposed Team MMSE precoding scheme largely outperforms both centralized and local precoding in all the considered scenarios. Perhaps surprisingly, the gains are significant even for a relatively small CSI sharing delay of 1 ms ($r = 0.99$, Figure 2b), a scenario where centralized precoding is still quite effective and outperforms local precoding. This suggests that the APs' local interference management capabilities based on timely local CSI should not be neglected in most practical scenarios. Indeed, our results suggests that significant rate gains can be achieved by a careful distributed precoding design, such as the proposed Team MMSE scheme, that is able to merge the benefits of local and centralized interference management.

Finally, Figure 2b shows that, for small CSI sharing delays, the aforementioned rate gains can be achieved by simple approximations of the Team MMSE scheme. In particular, in the considered scenario, we observe that the proposed structure-aware distributed precoding scheme is able to exploit the benefits of local and centralized interference management, at a significantly lower computational cost than the Team MMSE scheme. However, Figure 2a shows that this may not be the case for higher CSI sharing delays, since the Team MMSE solution shows non-negligible performance gains.

REFERENCES

[1] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021.

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

[3] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization," vol. 67, no. 10, pp. 6796–6815, Oct. 2019.

[4] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. and Netw.*, vol. 2019, no. 1, pp. 197, 2019.

[5] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, and Z. Zhang, "Cell-free massive MIMO systems with low resolution ADCs," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6844–6857, 2019.

[6] R. Nikbakht, R. Mosayebi, and A. Lozano, "Uplink fractional power control and downlink power allocation for cell-free networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 774–777, 2020.

[7] 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.

[8] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "Subspace-based pilot decontamination in user-centric scalable cell-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 4117–4131, 2023.

[9] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5022–5032, 2015.

[10] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes," *IEEE Trans. Commun.*, vol. 69, no. 11, Nov. 2021.

[11] L. Miretti, E. Björnson, and D. Gesbert, "Team MMSE precoding with applications to cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6242–6255, Aug. 2022.

[12] I. Atzeni, B. Gouda, and A. Tölli, "Distributed precoding design via over-the-air signaling for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1201–1216, Feb. 2021.

[13] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338–351, 2013.

[14] J. Zheng, J. Zhang, E. Björnson, and B. Ai, "Impact of channel aging on cell-free massive MIMO over spatially correlated channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6451–6466, 2021.

[15] W. Jiang and H. D. Schotten, "Deep learning-aided delay-tolerant zero-forcing precoding in cell-free massive MIMO," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–5.

[16] L. Miretti, R. L. G. Cavalcante, E. Björnson, and S. Stańczak, "UL-DL duality for cell-free massive MIMO with per-AP power and information constraints," *arXiv:2301.06520*, 2023.

[17] S. Yüksel and T. Başar, *Stochastic networked control systems: Stabilization and optimization under information constraints*, Springer Science & Business Media, 2013.

[18] L. Miretti, R. L. G. Cavalcante, and S. Stańczak, "Joint optimal beamforming and power control in cell-free massive MIMO," *Proc. IEEE Global Conf. Communications (GLOBECOM)*, 2022.

[19] A. Bazco-Nogueras, P. De Kerret, D. Gesbert, and N. Gresset, "Asymptotically achieving centralized rate on the decentralized network MISO channel," *IEEE Trans. Info. Theory*, vol. 68, no. 1, pp. 248–271, 2022.