Pseudo-Inverse vs Generalized Inverse for C-RAN Downlink Precoding

Mohammad M. Mojahedian, *Member, IEEE*, Reza Mosayebi, *Member, IEEE*, and Angel Lozano, *Fellow, IEEE*

***Abstract*—This paper tackles the problem of zero-forcing (ZF) precoding for the downlink of centralized radio-access networks operating in a cell-free fashion. While the customary workhorse of ZF precoding is the channel pseudo-inverse, because of the separate power constraint at each participating access point, the pseudo-inverse is not optimum. Rather, it can be improved upon by other inverses that allow conveying stronger signals (or, equivalently, consuming less power) while respecting the same ZF conditions. Motivated by the enormous disparity in computational cost between simple pseudo-inversion and general inversion, we ascertain the performance disadvantage of the former in a wide range of conditions. The welcome conclusion is that pseudo-inversion is close-to-optimum for all operational regimes of interest.**

I. Introduction

In a cellular network, each user is served by a single access point (AP) and the transmissions from other APs constitute interference. Owing to the invariance with the cell size of the signal-to-interference ratio (SIR), cellular networks have sustainedly increased their area capacity through densification [1]. This SIR invariance, however, is bound to break down once network densities are such that line-of-sight propagation comes to dominate and interference surges [2]. In conjunction with the ongoing transformation to software-defined networks, this motivates the new paradigm of centralized, possibly cloud-based, radio access networks (C-RANs) [3]. There, the AP transmitters consist solely of antennas, amplifiers, and upconverters, connected by powerful fronthaul lines to an edge datacenter hosting the baseband processing.

By default, C-RANs are cell-free structures where every AP potentially serves every user, inheriting and taking to the limit the principles of cell cooperation [4]. Influenced by massive MIMO, much of the cell-free literature posits conjugate-beamforming downlink precoders [5]–[7]. However, the potential of C-RANs emerges in full when their centralized nature is exploited to feature more sophisticated precoders. In particular, a linear zero-forcing (ZF) precoder should perform decidedly better than conjugate beamforming because:

- Interference then becomes a prime consideration, as opposed to being disregarded, which markedly improves the signal-to-interference-plus-noise ratios (SINRs).
- Thanks to these SINR improvements, much higher user loads can be tolerated, with the subsequent increases in spectral efficiency.
- While nulling interference, ZF precoders can keep the fading of the intended signals to a minimum; this offers the possibility of pilot-free downlink operation even in the absence of channel hardening—which, in cell-free settings, is only partial [8], [9]

Pioneering works do confirm the effectiveness of ZF precoding based on channel pseudo-inversion [6], [10]. However, while pseudo-inversion would be the optimum form for ZF in the face of a sum power constraint across all APs, better inverse forms exist in general for the actual per-AP power constraints. This is the thrust of the present paper, where we set out to establish the degree to which pseudo-inversion is suboptimum within the confines of ZF precoding.

II. Network and channel models

We consider networks featuring N APs and K users, all equipped with a single omnidirectional antenna. Some symbols are reserved for pilot transmissions from the users, based on which the channels are estimated by the APs. The remaining symbols are available for data transmission.

A. Large-scale Modeling

Distance-dependent pathloss with exponent η, in conjunction with shadowing, gives rise to a large-scale gain $G_{k,n}$ between the nth AP and the kth user. The corresponding downlink and uplink large-scale SNR equals $\mathsf{SNR}_{k,n} = G_{k,n}P_{\mathrm{t}}/\sigma^2$ with P_{t} the transmit power and σ^2 the noise power. Although P/σ^2 is taken as equal for uplink and downlink, any asymmetry could be readily absorbed into the number of pilot symbols (see Sec. III-A).

B. Small-scale Modeling

Besides $G_{k,n}$, the channel between the nth AP and the kth user features a small-scale fading coefficient $h_{k,n} \sim \mathcal{N}_{\mathbb{C}}(0,1)$, independent across APs and users.

For any snapshot of the large-scale parameters, under the premise that channel estimation errors and interference are treated by the decoder as additional Gaussian noise, user k can achieve a spectral efficiency of

$$C_k = \log_2(1 + \mathsf{sinr}_k), \tag{1}$$

where sinr_k denotes the SINR of user k. Adding (1) over all K users, we obtain the sum spectral efficiency. The spectral efficiencies in this paper are *gross*, meaning that pilot overheads are yet to be subtracted out.

C. Simulation Environment

For simulation purposes, we resort to a wrapped-around universe with $N = 100$ APs elevated 2 m above the users to avoid distance singularities. Under the premise of AP positions agnostic to the radio propagation, shadow fading renders the network approximately Poisson-like from the vantage of any user [11]. This approximation sharpens as the shadowing strengthens, being precise for relevant values thereof [12], [13]. Relying on this result, the AP positions are drawn uniformly at random. Likewise, the user positions are uniformly random.

Less otherwise stated, P_{t}/σ^2 such that $\mathsf{SNR}_{k,n} = 25$ dB at a distance d, where d would be the inter-AP spacing if the network were a hexagonal grid with the same spatial density. Under reasonable values for P_{t}, the bandwidth, and the pathloss intercept [14], this is compatible with ultradense deployments ($d \approx$ 10–20 m) and it would ensure interference-limited conditions should the network operate in cellular mode.

III. Channel Estimation and Data Transmission

A. Uplink Channel Estimation

Let N_{p} be the number of uplink symbols reserved for pilot transmissions on every fading coherence block, with every user being allocated N_{p}/K of those pilots. Disregarding pilot contamination, the MMSE fading estimate $\hat{h}_{k,n}$ gathered by the network upon observation at the nth AP of the N_{p}/K pilots emitted by user k satisfies $h_{k,n} = \hat{h}_{k,n} + \tilde{h}_{k,n}$ where [15, sec. 4.8]

$$\mathbb{E}\big[|\hat{h}_{k,n}|^2\big] = \frac{\frac{N_{\mathrm{p}}}{K}\mathsf{SNR}_{k,n}}{1+\frac{N_{\mathrm{p}}}{K}\mathsf{SNR}_{k,n}} \tag{2}$$

while

$$\tilde{h}_{k,n} \sim \mathcal{N}_{\mathbb{C}}\left(0, \frac{1}{1+\frac{N_{\mathrm{p}}}{K}\mathsf{SNR}_{k,n}}\right) \tag{3}$$

is uncorrelated error. The overhead corresponding to all the symbols consumed by pilots must be discounted from the gross spectral efficiency.

B. Downlink Data Transmission

Let $\boldsymbol{C}$ be the channel matrix that combines large- and small-scale components, i.e., whose (k,n)th entry is

$$[\boldsymbol{C}]_{k,n} = \sqrt{G_{k,n}}h_{k,n} \tag{4}$$

such that the column vector of observations at the K users is

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{v}, \tag{5}$$

where $\boldsymbol{v} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The vector of signals transmitted by the N APs is

$$\boldsymbol{x} = \sqrt{P_{\mathrm{t}}}\boldsymbol{T}\boldsymbol{s} \tag{6}$$

with $\boldsymbol{s}$ a vector containing the unit-power symbols intended for the K users and with $\boldsymbol{T}$ the $N \times K$ precoder. For the power constraints to be satisfied at the APs, it must hold that

$$\mathbb{E}\big[\boldsymbol{T}\boldsymbol{T}^*\big]_{n,n} \leq 1 \qquad n = 1, \ldots, N \tag{7}$$

where the expectation is over the distribution induced by the small-scale fading, for any given large-scale gains.

We hasten to emphasize that, because no downlink pilots are considered, users are not privy to $\boldsymbol{C}$ or to $\boldsymbol{T}$.

IV. ZF Precoding

The aim of ZF precoding is to eliminate the interference, and in essence this entails the inversion of $\boldsymbol{C}$. Broadly speaking, the applicable instrument is the generalized inversion, which subsumes any matrix $\boldsymbol{C}^-$ satisfying [16], [17]

$$\boldsymbol{C}\boldsymbol{C}^-\boldsymbol{C} = \boldsymbol{C}. \tag{8}$$

Chief among the generalized inverses stands the pseudo-inverse, $\boldsymbol{C}^\dagger = \boldsymbol{C}^*(\boldsymbol{C}\boldsymbol{C}^*)^{-1}$, which exhibits the smallest Frobenius norm among all generalized inverses. And, as it turns out, any other inverse relates to $\boldsymbol{C}^\dagger$ via

$$\boldsymbol{C}^- = \boldsymbol{C}^\dagger + \big(\boldsymbol{I} - \boldsymbol{C}^\dagger\boldsymbol{C}\big)\boldsymbol{A}, \tag{9}$$

where $\boldsymbol{A}$ is an arbitrary matrix and $(\boldsymbol{I} - \boldsymbol{C}^\dagger\boldsymbol{C})$ projects it onto the null space of $\boldsymbol{C}$.

Suppose for starters that the network has perfect knowledge of $\boldsymbol{C}$, which amounts to letting $N_{\mathrm{p}} \to \infty$. From (8), precoding directly with $\boldsymbol{T} = \boldsymbol{C}^-$ removes the interference completely and equalizes the signal powers at the users. This latter part is in fact an unnecessary imposition, and hence the precoder can retain its ZF nature while adopting the less restrictive form

$$\boldsymbol{T} = \boldsymbol{C}^-\mathrm{diag}\big(\sqrt{p_1}, \ldots, \sqrt{p_K}\big), \tag{10}$$

with p_k the signal power to be received by user k, normalized by P_{t}. This gives, at the users,

$$\boldsymbol{y} = \sqrt{P_{\mathrm{t}}}\,\mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K})\,\boldsymbol{s} + \boldsymbol{v}. \tag{11}$$

Without downlink pilots, as mentioned, users are unable to estimate $p_1, \ldots, p_K$. Therefore, $p_1, \ldots, p_K$ are not allowed to depend on the small-scale fading coefficients within $\boldsymbol{C}$, which are unknown to the users and cannot be estimated either, but only on the large-scale gains, which are known. Under that premise, the kth user enjoys

$$\mathsf{snr}_k^{\mathsf{ZF}} = \frac{P_{\mathrm{t}}}{\sigma^2}\,p_k. \tag{12}$$

Now letting N_{p} be finite, the network is no longer privy to $\boldsymbol{C}$, but only to its estimate $\hat{\boldsymbol{C}}$, and the ZF precoder becomes

$$\boldsymbol{T} = \hat{\boldsymbol{C}}^-\mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K}) \tag{13}$$

such that

$$\begin{aligned}\boldsymbol{y} &= \sqrt{P_{\mathrm{t}}}\,\big(\hat{\boldsymbol{C}} + \tilde{\boldsymbol{C}}\big)\hat{\boldsymbol{C}}^-\mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K})\,\boldsymbol{s} + \boldsymbol{v} &(14)\\ &= \sqrt{P_{\mathrm{t}}}\,\mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K})\,\boldsymbol{s}\\ &\quad + \underbrace{\sqrt{P_{\mathrm{t}}}\,\tilde{\boldsymbol{C}}\hat{\boldsymbol{C}}^-\,\mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K})\,\boldsymbol{s}}_{\text{Interference}} + \boldsymbol{v}. &(15)\end{aligned}$$

With interference now leaking because of imperfect channel estimation, and with $\tilde{\boldsymbol{C}}$ and $\hat{\boldsymbol{C}}^-$ unknown to the users,

$$\mathsf{sinr}_k^{\mathsf{ZF}} = \frac{p_k}{\mathbb{E}\big[\tilde{\boldsymbol{c}}_k\hat{\boldsymbol{C}}^-\mathrm{diag}(p_1, \ldots, p_K)\hat{\boldsymbol{C}}^{-*}\tilde{\boldsymbol{c}}_k^*\big] + \sigma^2/P_{\mathrm{t}}}, \tag{16}$$

where $\tilde{\boldsymbol{c}}_k$ is the kth row of $\tilde{\boldsymbol{C}}$, satisfying

$$\mathbb{E}\big[\tilde{\boldsymbol{c}}_k^* \tilde{\boldsymbol{c}}_k\big] = \mathrm{diag}\left(\frac{G_{k,1}}{1+\frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,1}}, \ldots, \frac{G_{k,N}}{1+\frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,N}}\right). \tag{17}$$

It follows that $\mathsf{sinr}_k^{\mathsf{ZF}}$ equals

$$\frac{P_\mathrm{t}}{\sigma^2} \frac{p_k}{\mathrm{Tr}\left(\mathbb{E}\big[\boldsymbol{T}\boldsymbol{T}^*\big]\,\mathrm{diag}\Big(\frac{\mathsf{SNR}_{k,1}}{1+\frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,1}}, \ldots, \frac{\mathsf{SNR}_{k,N}}{1+\frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,N}}\Big)\right)+1} \tag{18}$$

and, for any objective function $f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big)$, the optimum ZF precoding problem amounts to

$$\begin{aligned} \max_{\hat{\boldsymbol{C}}^-, p_1, \ldots, p_K} \quad & f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big) \\ \text{s.t.} \quad & \mathbb{E}\Big[\hat{\boldsymbol{C}}^- \mathrm{diag}(p_1, \ldots, p_K)\, \hat{\boldsymbol{C}}^{-*}\Big]_{n,n} \leq 1 \quad \forall n. \end{aligned} \tag{19}$$

This is in general a difficult, often nonconvex optimization, which has long motivated the interest in simpler alternatives revolving around the readily computable pseudo-inverse. The customary alternative to the optimum ZF precoder is then

$$\boldsymbol{T} = \hat{\boldsymbol{C}}^\dagger \mathrm{diag}\big(\sqrt{p_1}, \ldots, \sqrt{p_K}\big), \tag{20}$$

which entails optimizing over $p_1, \ldots, p_K$ only, namely

$$\begin{aligned} \max_{p_1, \ldots, p_K} \quad & f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big) \\ \text{s.t.} \quad & \mathbb{E}\Big[\hat{\boldsymbol{C}}^\dagger \mathrm{diag}\big(p_1, \ldots, p_K\big) \hat{\boldsymbol{C}}^{\dagger *}\Big]_{n,n} \leq 1 \quad \forall n \end{aligned} \tag{21}$$

with $\mathsf{sinr}_k^{\mathsf{ZF}}$ still given by (18), only with $\boldsymbol{T}$ in (20). Since the Frobenius norm of $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}^*]$ measures the sum transmit power and the pseudo-inverse exhibits the smallest such norm, (20) would embody the optimum ZF precoder under a sum power constraint. However, the per-AP constraints void that optimality and force a downscaling of $p_1, \ldots, p_K$, hence of the SINRs, and our goal is precisely to gauge the extent to which the solution to (21) is suboptimum relative to (19).

V. Pseudo-Inverse vs Optimum ZF

Inspired by [17], we circumvent the difficulty of solving (19) by means of the upper bound obtained by relaxing the per-AP power constraints (7) into the sum power constraint

$$\mathbb{E}\Big[\mathrm{Tr}\big(\boldsymbol{T}\boldsymbol{T}^*\big)\Big] \leq N. \tag{22}$$

Thus relaxed, (19) morphs into

$$\begin{aligned} \max_{\hat{\boldsymbol{C}}^-, p_1, \ldots, p_K} \quad & f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big) \\ \text{s.t.} \quad & \mathbb{E}\Big[\mathrm{Tr}\Big(\hat{\boldsymbol{C}}^- \mathrm{diag}(p_1, \ldots, p_K)\, \hat{\boldsymbol{C}}^{-*}\Big)\Big] \leq N \end{aligned} \tag{23}$$

and the gap between the solutions to (21) and (23) brackets the shortfall of pseudo-inversion relative to the optimum ZF precoder. Details on how to solve these respective optimizations are provided in Appendix A.

To ensure the broadest possible scope for the assessment of this gap, we consider the two extremes in terms of objective function:

- Maximum performance fairness across users, disregarding the aggregate. This corresponds to
$$f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big) = \min_k \mathsf{sinr}_k^{\mathsf{ZF}}. \tag{24}$$
- Maximum sum performance, disregarding fairness across users. Recalling (1), this is well represented by the sum spectral efficiency
$$f\big(\mathsf{sinr}_1^{\mathsf{ZF}}, \ldots, \mathsf{sinr}_K^{\mathsf{ZF}}\big) = \sum_{k=1}^K \log_2\big(1+\mathsf{sinr}_k^{\mathsf{ZF}}\big). \tag{25}$$

The load K/N is set to 0.8, far above the values that are typical with conjugate beamforming—this is the main motivation for ZF—and only slightly below the maximum possible one; this load $K/N = 0.8$ is indeed a desirable operating point for a C-RAN.

With all the pieces in place, we can proceed to assess the gap. Shown in Figs. 1 and 2 are, respectively for the min SINR and the sum spectral efficiency objectives, the gaps between the solutions to (21) and (23) in the form of CDFs taken over the large-scale gains. For each objective, various values of N_p are entertained, namely the baseline $N_\mathrm{p} = K$ (one pilot per user and coherence block), then $N_\mathrm{p} = 10K$, and finally $N_\mathrm{p} \to \infty$ (perfect channel estimation at the APs). For completeness, conjugate-beamforming performance curves targeting the same objectives are also included [5], [18]. Altogether, the following can be observed:

- ZF precoding is exceedingly superior to conjugate beamforming already with minimum pilot overhead ($N_\mathrm{p} = K$), and even more so for growing N_p.
- The gap between pseudo-inversion and optimum ZF is always small, and outright negligible in practically relevant conditions ($N_\mathrm{p} \lesssim 10K$).

Reinforcing the second observation, recall that the far edge of the gap is not the optimum ZF itself, but a bound to it. To gauge the tightness of this upper bound, Fig. 3 depicts the CDF of the corresponding per-AP transmit power relative to its maximum value, P_t, for the case $N_\mathrm{p} \to \infty$. At any given time, about half the APs operate—by as much as 5 dB—above P_t, suggesting that the upper bound might not be tight and bolstering the second observation.

The gap between pseudo-inversion and optimum ZF remains relatively stable for other ratios K/N. Its dependence with P_t/σ^2, in turn, is characterized in Figs. 6 and 7 for the min SINR objective. Examined both at the lower tail and at the median, the gap remains relatively very small except when the SNR at distance d ceases to be high while N_p is large; this would correspond to situations where ZF—by definition an inherently high-SNR strategy—is altogether inappropriate while the pilot overhead is unappealingly high.

VI. Precoded Downlink Pilots

Let us now assess whether the insight from the previous section changes if the network is equipped with downlink pilots. Hence herein it is assumed that user k is aware of precoded channel from the N APs. Then, $p_1, \ldots, p_K$ are allowed to depend on the small-scale fading. In the following subsections, two cases of perfect and imperfect channel estimation in users are examined.

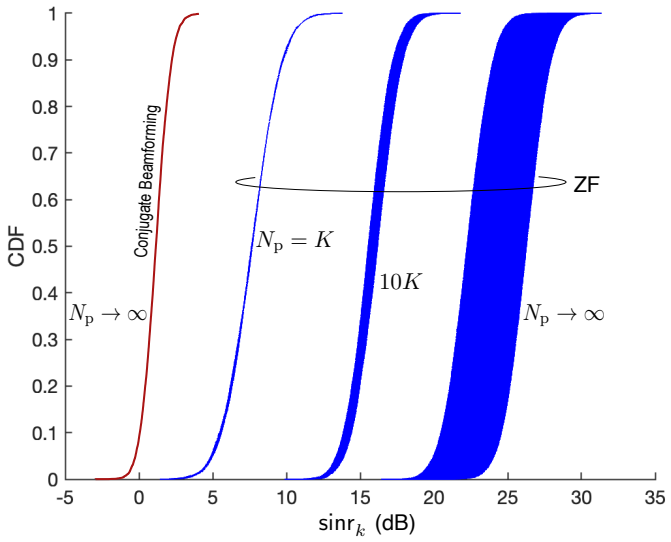

Fig. 1. CDF of the max-min SINR for $N_\mathrm{p} = K$, $N_\mathrm{p} = 10K$, and $N_\mathrm{p} \to \infty$. For each case, the respective shaded region extends from the pseudo-inverse performance to the optimum ZF upper bound. Also shown is the conjugate beamforming performance for $N_\mathrm{p} \to \infty$.

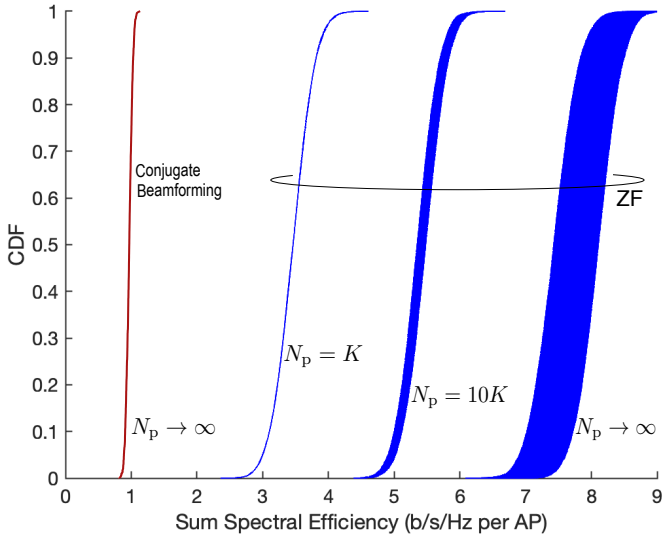

Fig. 2. CDF of the maximum gross sum spectral efficiency for $N_\mathrm{p} = K$, $N_\mathrm{p} = 10K$, and $N_\mathrm{p} \to \infty$. For each case, the respective shaded region extends from the pseudo-inverse performance to the optimum ZF upper bound. Also shown is the conjugate beamforming performance for $N_\mathrm{p} \to \infty$.

A. Users with Perfect Channel Estimation

Consider the case of perfect channel estimation. Then, (12) applies, only with p_k now allowed to depend on the small-scale fading. The per-AP power constraints become

$$\sum_{k=1}^{K} \mathbb{E}\Big[p_k \left|c_{n,k}^\dagger\right|^2\Big] \le 1 \qquad n = 1, \ldots, N. \tag{26}$$

So in the simplest case where we have an exact estimate of the channel, we want to solve the following optimization problem.

$$\max_{p_1,\ldots,p_K} \quad \min_k \frac{P_\mathrm{t}}{\sigma^2}\mathbb{E}[p_k] \tag{27}$$
$$\text{s.t.} \quad \sum_{k=1}^{K} \mathbb{E}\Big[p_k \left|c_{n,k}^\dagger\right|^2\Big] \le 1 \qquad n = 1, \ldots, N.$$

And as an upper-bound, per-AP power constraint can be replaced with sum-power constraint as

$$\max_{p_1,\ldots,p_K} \quad \min_k \frac{P_\mathrm{t}}{\sigma^2}\mathbb{E}[p_k] \tag{28}$$
$$\text{s.t.} \quad \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\Big[p_k \left|c_{n,k}^\dagger\right|^2\Big] \le N.$$

Details on how to solve the optimization problem (27) are provided in Appendix B.

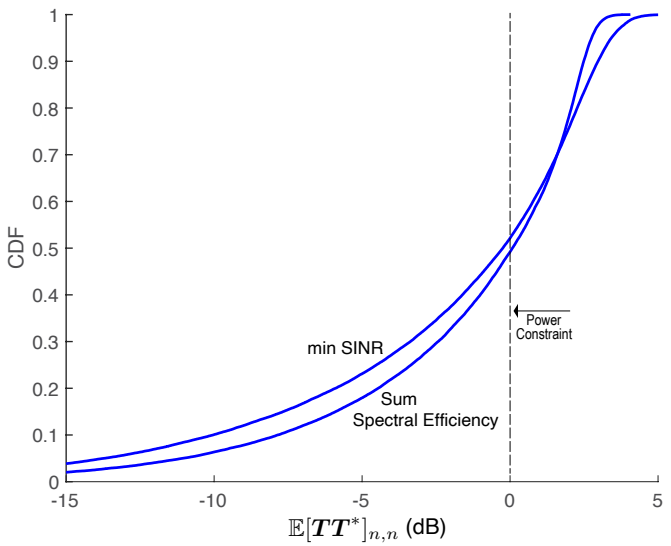

Fig. 3. CDF of the per-AP transmit power (relative to P_t) that attains the upper bound to optimum ZF precoding for $N_\mathrm{p} \to \infty$.

B. Users with Imperfect Channel Estimation

Now let us assume that we have downlink pilots and users do not have access to perfect channel estimates.

Replacing (13) in (5), the received vector can be represented as

$$\boldsymbol{y} = \sqrt{P_\mathrm{t}}\boldsymbol{C}\hat{\boldsymbol{C}}^\dagger \mathrm{diag}\big(\sqrt{p_1}, \ldots, \sqrt{p_K}\big)\boldsymbol{s} + \boldsymbol{v}, \tag{29}$$

and consequently the signal received by the kth user is

$$\begin{aligned} y_k &= \sqrt{P_\mathrm{t}}\boldsymbol{c}_k\hat{\boldsymbol{C}}^\dagger \mathrm{diag}\big(\sqrt{p_1}, \ldots, \sqrt{p_K}\big)\boldsymbol{s} + v_k \\ &= \sqrt{P_\mathrm{t}}\sqrt{p_k}\boldsymbol{c}_k\hat{\boldsymbol{c}}_k^\dagger s_k + \sum_{\ell\neq k}\sqrt{P_\mathrm{t}}\sqrt{p_\ell}\boldsymbol{c}_k\hat{\boldsymbol{c}}_\ell^\dagger s_\ell + v_k \\ &= \sqrt{P_\mathrm{t}}a_{kk}s_k + \sum_{\ell\neq k}\sqrt{P_\mathrm{t}}a_{k\ell}s_\ell + v_k, \end{aligned} \tag{30}$$

where $a_{kk} = \sqrt{p_k}\boldsymbol{c}_k\hat{\boldsymbol{c}}_k^\dagger$ and $a_{k\ell} = \sqrt{p_\ell}\boldsymbol{c}_k\hat{\boldsymbol{c}}_\ell^\dagger$. Now the goal is to estimate the desired signal coefficient a_{kk} using downlink pilots. To do this, consider a sequence of pilots $\sqrt{N_\mathrm{d}}\boldsymbol{\phi}_1, \sqrt{N_\mathrm{d}}\boldsymbol{\phi}_2, \ldots, \sqrt{N_\mathrm{d}}\boldsymbol{\phi}_K$, each with length N_d such

that $\|\boldsymbol{\phi}_k\| = 1$. Moreover, we assume that $\boldsymbol{\phi}_k$'s are orthonormal which makes it necessary that $N_{\mathrm{d}} \geq K$. By sending pilots, the transmitted signal is

$$\boldsymbol{x} = \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{T} \begin{bmatrix} \boldsymbol{\phi}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\phi}_K^{\mathrm{T}} \end{bmatrix}, \tag{31}$$

which results in the following received signal

$$\boldsymbol{y} = \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{C} \boldsymbol{T} \begin{bmatrix} \boldsymbol{\phi}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\phi}_K^{\mathrm{T}} \end{bmatrix} + \boldsymbol{v}, \tag{32}$$

and subsequently the following $1 \times N_{\mathrm{d}}$ received vector at user k.

$$\boldsymbol{y}_k = \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{c}_k \boldsymbol{T} \begin{bmatrix} \boldsymbol{\phi}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\phi}_K^{\mathrm{T}} \end{bmatrix} + \boldsymbol{v}_k \tag{33}$$

The kth user by projecting its received signal on the pilot $\boldsymbol{\phi}_k$ will have

$$\begin{aligned} \boldsymbol{y}_k \boldsymbol{\phi}_k &= \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{c}_k \boldsymbol{T} \begin{bmatrix} \boldsymbol{\phi}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\phi}_K^{\mathrm{T}} \end{bmatrix} \boldsymbol{\phi}_k + \boldsymbol{v}_k \boldsymbol{\phi}_k \\ &= \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{c}_k \boldsymbol{T} \boldsymbol{e}_k + \boldsymbol{v}_k \boldsymbol{\phi}_k \\ &= \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{c}_k \boldsymbol{t}_k + \boldsymbol{v}_k \boldsymbol{\phi}_k \\ &= \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} \boldsymbol{c}_k \hat{\boldsymbol{c}}_k^{\dagger} \sqrt{p_k} + \boldsymbol{v}_k \boldsymbol{\phi}_k \\ &= \sqrt{N_{\mathrm{d}} P_{\mathrm{t}}} a_{kk} + \boldsymbol{v}_k \boldsymbol{\phi}_k, \end{aligned} \tag{34}$$

where $\boldsymbol{e}_k$ denotes the vector with a 1 in the kth entry and 0 elsewhere. Therefore, the desired signal coefficient could be properly approximated as

$$\hat{a}_{kk} = \frac{1}{\sqrt{N_{\mathrm{d}} P_{\mathrm{t}}}} \boldsymbol{y}_k \boldsymbol{\phi}_k, \tag{35}$$

which has the following variance of estimation error.

$$\begin{aligned} \mathbb{E}\left[|a_{kk} - \hat{a}_{kk}|^2\right] &= \mathbb{E}\left[|\frac{1}{\sqrt{N_{\mathrm{d}} P_{\mathrm{t}}}} \boldsymbol{v}_k \boldsymbol{\phi}_k|^2\right] \\ &= \frac{\sigma^2}{N_{\mathrm{d}} P_{\mathrm{t}}}. \end{aligned} \tag{36}$$

With this estimation, the received signal y_k in (30) could be rewritten as

$$\begin{aligned} y_k = \sqrt{P_{\mathrm{t}}} \hat{a}_{kk} s_k + \sqrt{P_{\mathrm{t}}} \left(a_{kk} - \hat{a}_{kk}\right) s_k \\ + \sum_{\ell \neq k} \sqrt{P_{\mathrm{t}}} a_{k\ell} s_\ell + v_k, \end{aligned} \tag{37}$$

which results in the following SINR at user k.

$$\begin{aligned} \mathsf{sinr}_k &= \frac{\hat{a}_{kk}^2}{\mathbb{E}\left[|a_{kk} - \hat{a}_{kk}|^2\right] + \sum_{\ell \neq k} |a_{k\ell}|^2 + \frac{\sigma^2}{P_{\mathrm{t}}}} \\ &= \frac{\hat{a}_{kk}^2}{\sum_{\ell \neq k} |a_{k\ell}|^2 + \frac{\sigma^2}{P_{\mathrm{t}}} \frac{N_{\mathrm{d}}+1}{N_{\mathrm{d}}}}. \end{aligned} \tag{38}$$

Depending on the fairness or sum-throughput criterion, the APs find the power coefficients p_k's by optimization problems (39) or (40), respectively, and then sends the signal $\boldsymbol{x}$.

$$\begin{aligned} \max_{p_1,\ldots,p_K} \quad & \min_k \frac{P_{\mathrm{t}}}{\sigma^2} \mathbb{E}\left[\frac{p_k}{1 + \sum_{\ell=1}^K \gamma_{k,\ell}\, p_\ell}\right] \\ \text{s.t.} \quad & \sum_{k=1}^K \mathbb{E}\left[p_k \left|\hat{c}_{n,k}^{\dagger}\right|^2\right] \leq 1 \qquad n = 1, \ldots, N. \end{aligned} \tag{39}$$

$$\begin{aligned} \max_{p_1,\ldots,p_K} \quad & \mathbb{E}\left[\log\left(1 + \frac{P_{\mathrm{t}}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^K \gamma_{k,\ell}\, p_\ell}\right)\right] \\ \text{s.t.} \quad & \sum_{k=1}^K \mathbb{E}\left[p_k \left|\hat{c}_{n,k}^{\dagger}\right|^2\right] \leq 1 \qquad n = 1, \ldots, N. \end{aligned} \tag{40}$$

The simulation results related to resulting SINR and sum rate from optimization problems (39) and (40) are shown in Figs. 10 and 11 respectively.

VII. Scalability

A network-wide precoder encompassing the entire C-RAN is not scalable and, in large deployments with thousands of APs and users, it is outright unfeasible. Moreover, a network-wide precoder is an unnecessary overkill because, due to pathloss and shadowing, only a small share of APs convey substantial power to user k and only a small share of other users suffer substantial interference from the transmission to user k. This suggests that, in a large network, the vast majority of channel entries should be disregarded in terms of precoding, and the estimation of those channel entries should be foregone altogether.

Let us consider the scalability of the precoder in terms of those aspects that are inherent to a C-RAN, namely (*i*) precoder computational cost, and (*ii*) channel estimation. The encoding and remaining pre-processing tasks are as in a cellular network, one chain per user, hence inherently scalable.

We measure the cost, denoted by M, by the number of complex multiply-and-accumulate (MA) operations accrued computing and applying the precoder coefficients. In turn, we denote by L the number of channel coefficients to be estimated. For growing N and K, we want $M/N = \mathcal{O}(1)$ and $L/N = \mathcal{O}(1)$ as in a cellular network.

Measured in MA operations, the cost of $N \times N$ matrix inversions or multiplications is $\mathcal{O}(N^3)$ while the multiplication of $N \times K$ and $K \times N$ matrices costs $\mathcal{O}(KN^2)$. Our goal here is not to present a detailed complexity analysis, which would require positing specific implementations, but rather to establish scalability. With this in mind, these measures suffice and simpler operations such as additions can be neglected, leading to the following considerations:

VIII. Sparse ZF Precoding

Even though, because of pathloss and shadowing, the channel matrix $\boldsymbol{C}$ has most of its mass concentrated on a small share of its entries, a network-wide ZF precoder requires estimates of every entry of $\boldsymbol{C}$ and then processes every such entry. The path to scalability lies precisely in recognizing and exploiting the nature of $\boldsymbol{C}$.

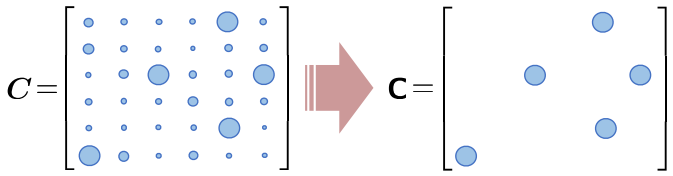

Fig. 4. Direct channel matrix sparsification.

An intuitive idea could be to zero out all but the dominant entries of $\boldsymbol{C}$, as in Fig. 4, thereby obtaining a sparse matrix $\mathbf{C}$ whose estimate would then be plugged into the various expressions in lieu of $\boldsymbol{C}$ itself. Unfortunately, such a direct sparsification might yield a channel matrix that is sparse, but unbalanced, with some users heavily favored by many connections while others are outright disconnected from the network. Likewise, some APs might be essentially taken out of service. The goal is therefore to generate a sparse channel matrix that is balanced across rows and columns, and to then capitalize on this sparsity to obtain a ZF precoder that is scalable. This objective is tackled in [19] for uplink MMSE reception, with the key to the derived solution being the delineation of a suitable subset of users to be processed by each AP and a suitable subset of APs to serve each user. We next tackle the challenge in the context of ZF precoding.

A. Formulation

Recalling (20), the pseudoinverse ZF precoder is

$$\boldsymbol{T} = \hat{\boldsymbol{C}}^* \big(\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}^*\big)^{-1} \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K}) \tag{41}$$

and, denoting by $\hat{\boldsymbol{c}}_n^*$ the nth row of $\hat{\boldsymbol{C}}^*$, the precoding vector at the nth AP is

$$[\boldsymbol{T}]_{n,:} = \hat{\boldsymbol{c}}_n^* \big(\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}^*\big)^{-1} \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K}). \tag{42}$$

Let us now restrict to a subset $\mathcal{N}_k$ the APs that observe uplink pilots from user k. Then, rather than $\hat{\boldsymbol{C}}$, the channel matrix estimate obtained by the network is $\hat{\mathbf{C}}$, defined as

$$[\hat{\mathbf{C}}]_{k,n} = \begin{cases} \sqrt{G_{k,n}}\, \hat{h}_{k,n} & n \in \mathcal{N}_k \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

Next, let us curtail to a subset $\mathcal{K}_n$ the users involved in producing the precoder at the nth AP, and let us identify the submatrix $\hat{\mathbf{C}}_n$ obtained by selecting those rows of $\hat{\mathbf{C}}$ that correspond to users in $\mathcal{K}_n$. From $\hat{\mathbf{C}}_n$, we can generate for the nth AP the modified precoding vector $[\mathbf{T}]_{n,:}$ with entries

$$[\mathbf{T}]_{n,k} = \begin{cases} \hat{\mathbf{c}}_n^* \big(\hat{\mathbf{C}}_n \hat{\mathbf{C}}_n^*\big)^{-1} \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K}) & k \in \mathcal{K}_n \\ 0 & \text{otherwise,} \end{cases} \tag{44}$$

where $\hat{\mathbf{c}}_n^*$ is the nth row of $\hat{\mathbf{C}}_n^*$. With all APs considered, $\mathbf{T}$ is the modified ZF precoder for the entire network and the users receive, in place of $\boldsymbol{y}$,

$$\mathbf{y} = \sqrt{P_\mathrm{t}} \boldsymbol{C}\mathbf{T}\boldsymbol{s} + \boldsymbol{v}. \tag{45}$$

The precoder $\mathbf{T}$, as well as the matrices $\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_N$ from which it derives, are all sparse as per the subsets $\mathcal{N}_1, \ldots, \mathcal{N}_K$ and $\mathcal{K}_1, \ldots, \mathcal{K}_N$. And, provided these subsets are delineated properly, the performance on the basis of $\mathbf{y}$ is close to the original one made possible by $\boldsymbol{y}$. Precisely, user k now enjoys

$$\begin{aligned} \mathsf{y}_k &= \sqrt{P_\mathrm{t}} \big(\hat{\boldsymbol{c}}_k + \tilde{\boldsymbol{c}}_k\big)\mathbf{T}\boldsymbol{s} + \boldsymbol{v} \\ &= \sqrt{P_\mathrm{t}} \hat{\boldsymbol{c}}_k \mathbf{t}_k s_k + \underbrace{\sqrt{P_\mathrm{t}} \sum_{\ell \neq k} \hat{\boldsymbol{c}}_k \mathbf{t}_\ell s_\ell}_{\mathrm{I}_1} \\ &\quad + \underbrace{\sqrt{P_\mathrm{t}} \sum_{\ell=1}^{K} \tilde{\boldsymbol{c}}_\ell \mathbf{t}_\ell s_\ell}_{\mathrm{I}_2} + \boldsymbol{v}. \end{aligned} \tag{46}$$

where $\mathbf{t}_k = [\mathbf{T}]_{:,k}$ is the kth column of $\mathbf{T}$. With interference now leaking because of the deviation of subset ZF precoder from pseudo-inverse ZF precoder (I_1) and imperfect channel estimation (I_2),

$$\mathsf{sinr}_k^{\mathsf{SZF}} = \frac{\big|\hat{\boldsymbol{c}}_k \mathbf{t}_k\big|^2}{\sum_{\ell \neq k} \big|\hat{\boldsymbol{c}}_k \mathbf{t}_\ell\big|^2 + \sum_\ell \big|\tilde{\boldsymbol{c}}_k \mathbf{t}_\ell\big|^2 + \sigma^2/P_\mathrm{t}}. \tag{47}$$

If we increase the uplink SNR or increase the N_p equivalently, the channel estimation error reduces and the I_2 interference expression disappears. Moreover, the closer the subset ZF precoder $\mathbf{T}$ gets to the ZF precoder $\boldsymbol{T}$ in (41), the smaller the I_1 interference term. Due to the pseudo-inverse continuity [20], if for all APs, $\hat{\mathbf{c}}_n^* \to \hat{\boldsymbol{c}}_n^*$, the subset ZF precoder $\mathbf{T}$ tends to $\boldsymbol{T}$. For the convergence of two random vectors $\hat{\mathbf{c}}_n^*$ and $\hat{\boldsymbol{c}}_n^*$, convergence in 2nd mean can be considered, which can be expanded as follows

$$\begin{aligned} \mathbb{E}\left[\|\hat{\mathbf{c}}_n^* - \hat{\boldsymbol{c}}_n^*\|^2\right] &= \mathbb{E}\left[\sum_{k \notin \mathcal{K}_n} G_{k,n} |\hat{h}_{k,n}|^2\right] \\ &= \sum_{k \notin \mathcal{K}_n} G_{k,n} \frac{\frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,n}}{1 + \frac{N_\mathrm{p}}{K}\mathsf{SNR}_{k,n}}. \end{aligned} \tag{48}$$

Roughly speaking, the smallness of (48) indicates the smallness of the interference term I_1. From (47) and (48), the following can be observed:

- Large N_p and high SNR: When N_p is large, the interference term I_2 caused by channel estimation error is negligible and the noise effect σ^2/P_t can be ignored as the SNR increases. Thus (47) and (48) can be approximated as

$$\mathsf{sinr}_k^{\mathsf{SZF}} \simeq \frac{\mathbb{E}\left[\big|\hat{\boldsymbol{c}}_k \mathbf{t}_k\big|^2\right]}{\sum_{\ell \neq k} \mathbb{E}\left[\big|\hat{\boldsymbol{c}}_k \mathbf{t}_\ell\big|^2\right]}. \tag{49}$$

$$\mathbb{E}\left[\|\hat{\mathbf{c}}_n^* - \hat{\boldsymbol{c}}_n^*\|^2\right] \simeq \sum_{k \notin \mathcal{K}_n} G_{k,n}. \tag{50}$$

  Therefore, to have a smaller amount of interference and thus a larger SINR, it is better for each AP to put the users in the set $\mathcal{K}_n$ who have the largest $G_{k,n}$.
- Large N_p and low SNR: In this case, noise is dominant and we have:

$$\mathsf{sinr}_k^{\mathsf{SZF}} \simeq \frac{P_\mathrm{t}}{\sigma^2} \mathbb{E}\left[\big|\hat{\boldsymbol{c}}_k \mathbf{t}_k\big|^2\right]. \tag{51}$$

B. Cost Analysis

The cost of obtaining the subset ZF precoder matrix in (44) satisfies

$$\frac{M_{\mathsf{rx}}}{N} = \mathcal{O}\left(\frac{\sum_{n=1}^{N}|\mathcal{K}_n|^{\nu}}{N}\right), \tag{52}$$

where is $\mathcal{O}(1)$ given that the sizes of the subsets $\mathcal{K}_1,\ldots,\mathcal{K}_N$ are sub-linear with respect to K and K/N is fixed. The following is the cost of the linear combining

$$\frac{M_{\mathsf{comb}}}{N} = \frac{\sum_{n=1}^{N}|\mathcal{K}_n|}{N}, \tag{53}$$

which is again equal to $\mathcal{O}(1)$ for fixed K/N. Finally, the number of channel coefficients required will be of order $\mathcal{O}(1)$ as shown below.

$$\frac{L}{N} = \frac{\sum_{k=1}^{K}|\mathcal{N}_k|}{N}. \tag{54}$$

C. User Selection Policy

Our desiderata for selection of users and constituting the sets $\mathcal{K}_1,\ldots,\mathcal{K}_N$ are

- Based on the large-scale coefficients to avoid having to update the sets $\mathcal{K}_1,\ldots,\mathcal{K}_N$ with rapid small-scale coefficients changes.
- Near-optimal performance
- Scalable,

Based on the above-mentioned criterion, the proposed policy for choosing $\mathcal{K}_n$ is to select $|\mathcal{K}_n|$ users with the largest $G_{n,k}$.

D. AP Subset Selection

In a way similar to user selection, we assume that $\mathcal{N}_k$ is the set of $|\mathcal{N}_k|$ APs with the largest $G_{n,k}$. Moreover, it is assumed that the nth AP estimates the channel coefficients corresponding to the users in $\mathcal{K}_n$ and therefore none of the vector $\hat{\mathbf{c}}_n^*$ elements are zero.

E. Evaluation

Consider a network with $N = 100$ APs and $K = 80$ users. The subset size $|\mathcal{K}_n| = 8$ is identical for all APs, while $\mathcal{N}_k$ contains the $\frac{N}{K}|\mathcal{K}_n|$ strongest large-scale channel coefficients between user k and APs. Having three values of $N_{\mathrm{p}} = K$, $N_{\mathrm{p}} = 10K$, and $N_{\mathrm{p}} \to \infty$, for the min SINR objective, upper and lower-bounds for median SINR are depicted as a function of P_{t}/σ^2 in Fig. 5.

By increasing the P_{t}/σ^2 to a value of about 20 dB, the median SINR's are almost equal for different cases. This is because in high SNRs, the dominant term at the denominator of the SINR relation in (47) is the I_1 interference term. This interference term for large SNRs will be a function of $\sum_{k\notin\mathcal{K}_n} G_{k,n}$, which depends only on the size of the subsets.

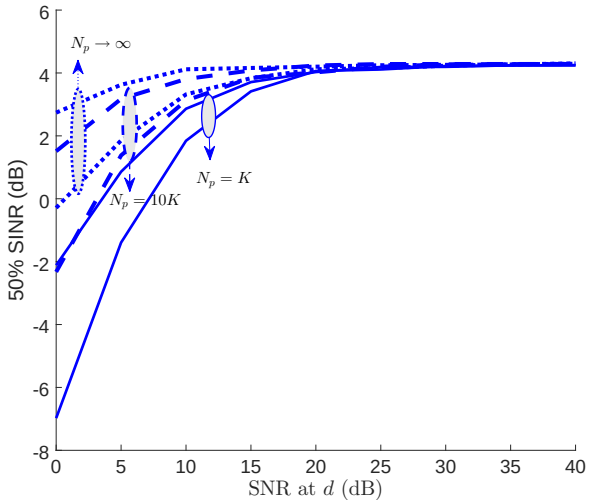

Fig. 5. 50% of the max-min SINR as a function of the SNR at distance d for $N_{\mathrm{p}} = K$ (solid), $N_{\mathrm{p}} = 10K$ (dashed), and $N_{\mathrm{p}} \to \infty$ (dotted). For each case, For each case, the upper- and lower-bounds of SZF performance are shown.

IX. Summary

The performance of the optimum ZF downlink precoder can be approached closely with channel pseudo-inversion, dodging the much more involved generalized inversion. This holds for performance objectives at both ends of the fairness axis, and for varying network loads. The only possible exception is a combination of low SNRs and very precise channel estimation, an uninteresting regime in terms of pilot overhead and, in fact, an uninteresting regime for ZF precoding in general.

The specific results presented in the paper correspond to a single-slope pathloss model, and the absolute values would surely change for a dual-slope pathloss, but there is no reason to think that the conclusions would.

Pilot contamination has been disregarded and, while it would be of interest to incorporate it, ultradense deployments do exhibit (because of their short range) very high frequency coherences and, most likely, also very high time coherences. The combined fading coherence is sure to be in the thousands of resource units, a very benevolent situation in terms of pilot contamination, even under random pilot assignment [7] and let alone with contamination mitigation procedures such as the ones propounded in [5, sec. IV] or in [21], [22].

Looking forward, it would be of interest to verify whether, also in the face of downlink precoded pilots, channel estimation at the users, and fading-dependent channel allocation, it holds that pseudo-inversion is near-optimum.

Acknowledgment

Work supported by the European Research Council under the H2020 Framework Programme/ERC grant 694974, and by MINECO's Projects RTI2018-102112 and RTI2018-101040.

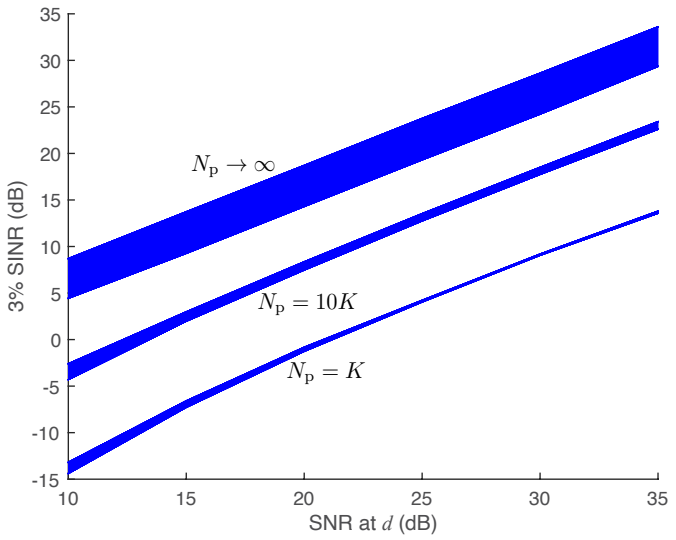

Fig. 6. 3% of the max-min SINR as a function of the SNR at distance d for $N_{\rm p} = K$, $N_{\rm p} = 10K$, and $N_{\rm p} \to \infty$. For each case, the respective shaded region extends from the pseudo-inverse performance to the optimum ZF upper bound.

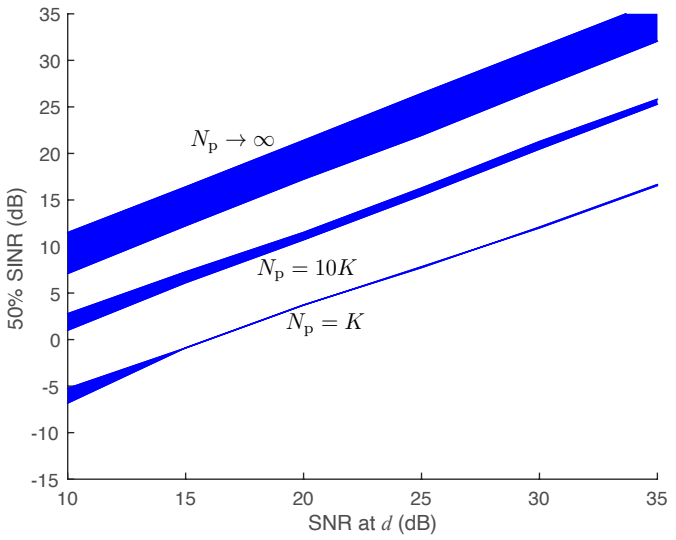

Fig. 7. 50% of the max-min SINR as a function of the SNR at distance d for $N_{\rm p} = K$, $N_{\rm p} = 10K$, and $N_{\rm p} \to \infty$. For each case, the respective shaded region extends from the pseudo-inverse performance to the optimum ZF upper bound.

Appendix A

Let us first consider the min-SINR objective. The per-AP power constraints in (21) can be rewritten as

$$\sum_{k=1}^{K} p_k \underbrace{\mathbb{E}\Big\{\big|\big[\hat{\boldsymbol{C}}^{\dagger}\big]_{n,k}\big|^2\Big\}}_{b_{n,k}} \leq 1 \qquad n = 1, \ldots, N$$

and, via the matrix $\boldsymbol{B} = [b_{n,k}]$ and $\boldsymbol{p} = \big[p_1, \ldots, p_K\big]^{\mathrm{T}}$, compactly as $\boldsymbol{B}\boldsymbol{p} \leq \boldsymbol{1}$ with the inequality being entry-wise and with $\boldsymbol{1}$ standing for the all-one vector. Similarly, the sum power constraint in (22) is equivalent to $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{p} \leq N$ where $\boldsymbol{b}$ is a column vector whose kth entry equals $\sum_{n=1}^{N} b_{n,k}$. Turning then to $\mathsf{sinr}_k^{\mathsf{ZF}}$, it is useful to rewrite (18) as

$$\mathsf{sinr}_k^{\mathsf{ZF}} = \frac{P_{\mathrm{t}}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}, \tag{55}$$

which is a linear-fractional function of $p_1, \ldots, p_K$ given

$$\gamma_{k,\ell} = \sum_{n=1}^{N} \frac{\mathsf{SNR}_{k,n}}{1 + \frac{N_{\mathrm{p}}}{K} \mathsf{SNR}_{k,n}}\, b_{n,\ell}. \tag{56}$$

Altogether, the pseudo-inverse precoding problem under a min SINR objective becomes

$$\begin{aligned} \max_{p_1,\ldots,p_K} \quad & \min_k \frac{P_{\mathrm{t}}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell} \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{p} \leq \boldsymbol{1}, \end{aligned} \tag{57}$$

which is a quasilinear optimization with linear constraints that can be solved through the bisection method. The corresponding upper bound in (23), subject to a sum power constraint, is achieved by pseudo-inversion and the ensuing optimization of $p_1, \ldots, p_K$ is again quasilinear with linear constraints.

Turning now to the sum spectral efficiency objective, from (55) it corresponds to

$$\begin{aligned} f(\mathsf{sinr}_k^{\mathsf{ZF}}) &= \sum_{k=1}^{K} \log_2\left(1 + \frac{P_{\mathrm{t}}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}\right) && (58) \\ &= \sum_{k=1}^{K} \log_2\left(1 + \frac{P_{\mathrm{t}}}{\sigma^2} p_k + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right) \\ &\quad - \sum_{k=1}^{K} \log_2\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right), && (59) \end{aligned}$$

which is difference of two concave functions of $p_1, \ldots, p_K$, hence in the scope of the difference-of-convex optimizations [23], [24]. The premise of this method is the linear approximation of the negative term around an initial solution through a first-order expansion, solving the resulting convex problem, and then updating the solution. Let $\boldsymbol{p}^{(0)}$ be an initial guess; the first-order approximation of the negative term around $\boldsymbol{p}^{(0)}$ is

$$\begin{aligned} \sum_{k=1}^{K} \log_2\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right) &\simeq \sum_{k=1}^{K} \log_2\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell^{(0)}\right) \\ &\quad + \nabla_{\boldsymbol{p}} \left(\sum_{k=1}^{K} \log_2\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right)\right)^{\mathrm{T}} \Bigg|_{\boldsymbol{p}=\boldsymbol{p}^{(0)}} \big(\boldsymbol{p} - \boldsymbol{p}^{(0)}\big), \end{aligned} \tag{60}$$

where the ith entry of $\nabla_{\boldsymbol{p}}(\cdot)$ equals

$$\frac{\partial \sum_{k=1}^{K} \log_2\big(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\big)}{\partial p_i} = \sum_{k=1}^{K} \frac{\gamma_{k,i} \log_2 e}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}. \tag{61}$$

Plugging (60) and (61) into (59), a concave function is obtained and denoted $f_0\big(\mathsf{sinr}_1^{\mathsf{ZF}},\ldots,\mathsf{sinr}_K^{\mathsf{ZF}}\big)$. Then, $\boldsymbol{p}^{(1)}$ can be found as

$$\begin{aligned}\boldsymbol{p}^{(1)} = \underset{p_1,\ldots,p_K}{\arg\max} \quad & f_0\big(\mathsf{sinr}_1^{\mathsf{ZF}},\ldots,\mathsf{sinr}_K^{\mathsf{ZF}}\big) \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{p} \le \mathbf{1}. \end{aligned} \tag{62}$$

Subsequently expanding $\boldsymbol{p}^{(1)}$, the process is repeated until the improvement in the objective is below 0.01 b/s/Hz. For the corresponding upper bound, the same procedure applies only with $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{p} \le N$ in (62). Various random choices for $\boldsymbol{p}^{(0)}$ are tested, including the waterfilling solution [15, sec. 5.3].

Appendix B

By the linearity of expectation, the per-AP power constraints in (27) can be rewritten as

$$\mathbb{E}\Big[\sum_{k=1}^{K} p_k \underbrace{\big|c_{n,k}^{\dagger}\big|^2}_{b_{n,k}}\Big] \le 1 \qquad n=1,\ldots,N$$

and, via the matrix $\boldsymbol{B} = [b_{n,k}]$ and $\boldsymbol{p} = \big[p_1,\ldots,p_K\big]^{\mathrm{T}}$, compactly as $\mathbb{E}\left[\boldsymbol{B}\boldsymbol{p}\right] \le \mathbf{1}$ with the inequality being entry-wise and with $\mathbf{1}$ standing for the all-one vector. Therefore, the optimization problem (27) can be rewritten as

$$\begin{aligned}\max_{p_1,\ldots,p_K} \quad & \min_k \frac{P_{\mathsf{t}}}{\sigma^2}\mathbb{E}\left[p_k\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\boldsymbol{B}\boldsymbol{p}\right] \le \mathbf{1}. \end{aligned} \tag{63}$$

Let $\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)},\ldots,\boldsymbol{B}^{(M)}$ be M samples drawn independently and identically due to the Gaussian distribution of small-scale fading coefficients. Then, let denote $\boldsymbol{p}^{(i)}$ as the solution of the following optimization problem

$$\begin{aligned}\boldsymbol{p}^{(i)} = \underset{\boldsymbol{p}}{\arg\max} \quad & \min_k \frac{P_{\mathsf{t}}}{\sigma^2} p_k \\ \text{s.t.} \quad & \boldsymbol{B}^{(i)}\boldsymbol{p} \le \mathbf{1}, \end{aligned} \tag{64}$$

which results in independent and identically distributed (IID) $\boldsymbol{p}^{(i)}$ for $i=1,2,\ldots,M$. Indeed, $\boldsymbol{p}^{(i)}$ is a function of $\boldsymbol{B}^{(i)}$ as $\boldsymbol{p}^{(i)} = f(\boldsymbol{B}^{(i)})$, and so iid $\boldsymbol{B}^{(i)}$'s result in iid $\boldsymbol{p}^{(i)}$'s. By utilizing the law of large numbers, the expectations in (63) could be estimated as

$$\mathbb{E}\left[p_k\right] \simeq f_M(p_k) = \frac{1}{M}\sum_{i=1}^{M} p_k^{(i)}, \tag{65}$$

$$\mathbb{E}\left[\boldsymbol{B}\boldsymbol{p}\right] \simeq \boldsymbol{g}_M(\boldsymbol{p}) = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{B}^{(i)}\boldsymbol{p}^{(i)}. \tag{66}$$

Both estimations are unbiased, i.e.,

$$\mathbb{E}\left[f_M(p_k)\right] = \mathbb{E}\left[p_k\right], \tag{67}$$

$$\mathbb{E}\left[\boldsymbol{g}_M(\boldsymbol{p})\right] = \mathbb{E}\left[\boldsymbol{B}\boldsymbol{p}\right]. \tag{68}$$

Moreover, the variance of estimations are

$$\begin{aligned}\mathrm{Var}\left[f_M(p_k)\right] &= \frac{1}{M^2}\sum_{i=1}^{M}\mathrm{Var}\left[p_k^{(i)}\right] \\ &= \frac{1}{M}\mathrm{Var}\left[p_k\right], \end{aligned} \tag{69}$$

where by assuming that the p_k's variance is bounded, it tends to zero as the number of samples M increases.

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{g}_M(\boldsymbol{p})\right] &= \frac{1}{M^2}\sum_{i=1}^{M}\mathrm{Var}\left[\boldsymbol{B}^{(i)}\boldsymbol{p}^{(i)}\right] \\ &= \frac{1}{M}\mathrm{Var}\left[\boldsymbol{B}\boldsymbol{p}\right] \le \frac{1}{M}\mathbf{1}, \end{aligned} \tag{70}$$

where $\mathrm{Var}\left[\cdot\right]$ and $\le$ operators are both element-wise.

Given the sample average functions defined in (65) and (66), we approximate the optimization problem (63) with the following

$$\begin{aligned}\mathsf{SNR}_M = \max_{\boldsymbol{p}^{(1)},\ldots,\boldsymbol{p}^{(M)}} \quad & \min_k \frac{P_{\mathsf{t}}}{\sigma^2} f_M(p_k) \\ \text{s.t.} \quad & \boldsymbol{g}_M(\boldsymbol{p}) \le \mathbf{1}. \end{aligned} \tag{71}$$

There are two points about (71).

- First, as M increases, sample average tends to expectation by law of large numbers, and SNR_∞ will actually be the exact solution to (63).
- Second, SNR_M is an increasing function of M.

To show the increasing nature of SNR_M, assume that $\boldsymbol{p}^{(1)},\ldots,\boldsymbol{p}^{(M)}$ are optimal solutions corresponding to the $\boldsymbol{B}^{(1)},\ldots,\boldsymbol{B}^{(M)}$. Then, it is straightforward to check that, for the IID set $\boldsymbol{B}^{(1)},\ldots,\boldsymbol{B}^{(M+1)}$, the power constraint is satisfied by $\boldsymbol{q}^{(1)},\ldots,\boldsymbol{q}^{(M+1)}$ such that

$$\boldsymbol{q}^{(i)} = \frac{M+1}{M}\boldsymbol{p}^{(i)}, \quad i=1,\ldots,M \tag{72}$$

$$\boldsymbol{q}^{(M+1)} = \mathbf{0}. \tag{73}$$

Therefore, we have

$$\mathsf{SNR}_{M+1} \ge \frac{P_{\mathsf{t}}}{\sigma^2}\frac{1}{M+1}\sum_{i=1}^{M+1}\boldsymbol{q}^{(i)} \tag{74}$$

$$= \frac{P_{\mathsf{t}}}{\sigma^2}\frac{1}{M+1}\sum_{i=1}^{M}\frac{M+1}{M}\boldsymbol{p}^{(i)} \tag{75}$$

$$= \mathsf{SNR}_M \tag{76}$$

Now let's get to the upper-bound in (28). By defining the vector $\boldsymbol{b} = [b_k]$ where

$$b_k = \sum_{n=1}^{N}\big|c_{n,k}^{\dagger}\big|^2, \tag{77}$$

and replacing the expectation with sampling average, we are going to solve the following optimization problem

$$\begin{aligned}\mathsf{SNR}_M^{\mathsf{U}} = \frac{P_{\mathsf{t}}}{\sigma^2}\max_{\boldsymbol{p}^{(1)},\ldots,\boldsymbol{p}^{(M)}} \quad & \min_k \frac{1}{M}\sum_{i=1}^{M} p_k^{(i)} \\ \text{s.t.} \quad & \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{b}^{(i)\mathrm{T}}\boldsymbol{p}^{(i)} \le N, \end{aligned} \tag{78}$$

where $\boldsymbol{b}^{(1)},\ldots,\boldsymbol{b}^{(M)}$ are M iid copies of the vector $\boldsymbol{b}$. Again, it is easy to show that $\mathsf{SNR}_M^{\mathsf{U}}$ is ascending with M.

For a scenario with $N=20$ APs and $K=16$ users, for three values of $M=1,5,10$, the CDF of SNR_M and $\mathsf{SNR}_M^{\mathsf{U}}$ are depicted in Fig. 8. As can be seen from the figure for both

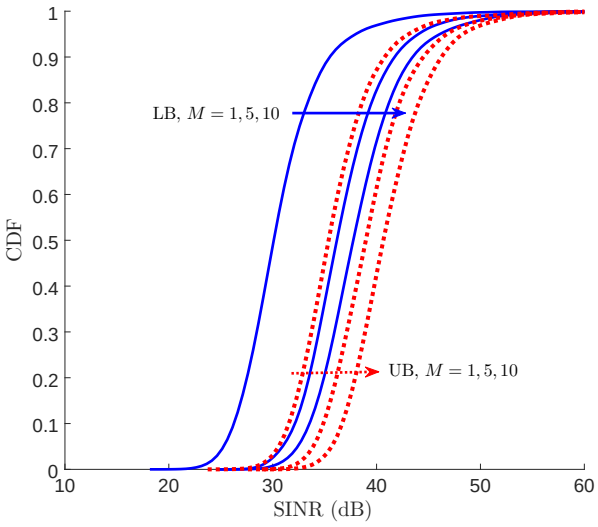

Fig. 8. Solid lines, CDF of the lower-bound SINR for $N_{\rm p}, N_{\rm d} \to \infty$ and three different values of $M = 1, 5, 10$ as the solution of stochastic optimization problem (71). Dashed lines, same quantities for upper-bound SNR obtained by (78).

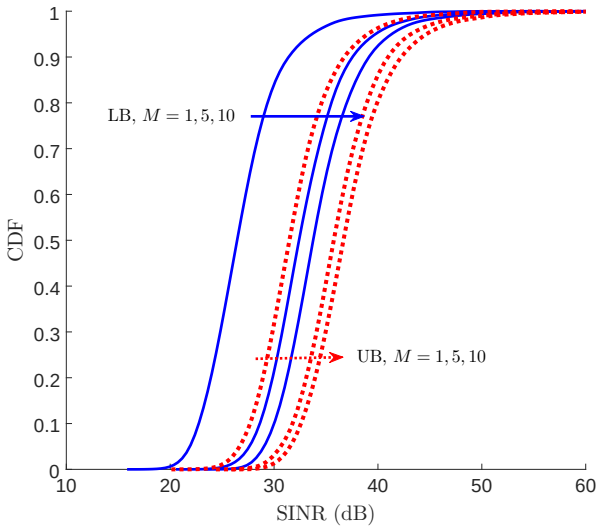

Fig. 10. Solid lines, CDF of the lower-bound SINR for $N_{\rm p} \to \infty$, $N_{\rm d} = K$ and three different values of $M = 1, 5, 10$ as the solution of stochastic optimization problem (71). Dashed lines, same quantities for upper-bound SNR obtained by (78).

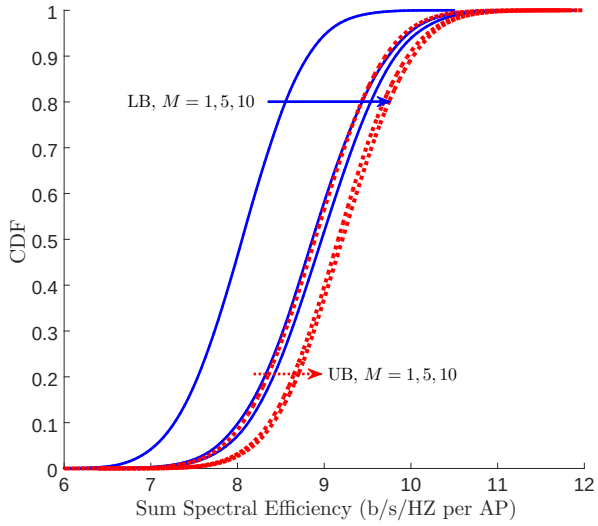

Fig. 9. CDF of the upper- and lower-bounds Sum spectral efficiency per AP for $N_{\rm p}, N_{\rm d} \to \infty$ and three different values of $M = 1, 5, 10$.

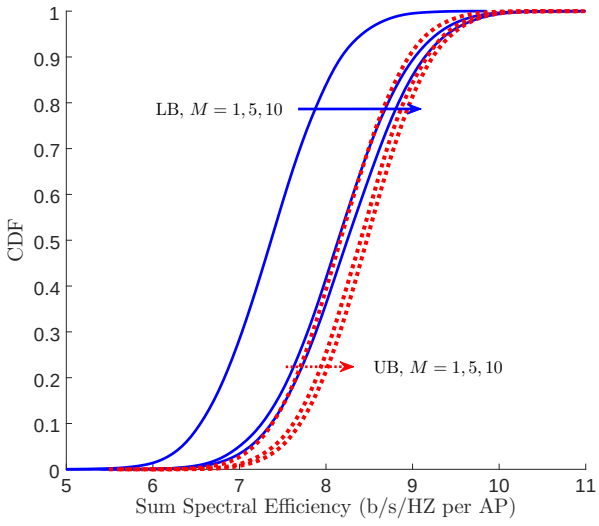

Fig. 11. Solid lines, CDF of the lower-bound pilot-assisted per-AP throughput per unit bandwidth for $N_{\rm p} \to \infty$, $N_{\rm p} = K$ and three different values of $M = 1, 5, 10$. Dashed lines, same quantities for upper-bound.

the upper- and lower- bounds, with increasing M, the curves become closer, indicating the convergence of the solutions of problems (71) and (78) as $M \to \infty$.

Similarly, for sum spectral efficiency, the following sample average function must be maximized according to the power constraint.

$$f_M(\boldsymbol{p}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} \log \left(1 + \frac{P_{\rm t}}{\sigma^2} p_k^{(i)} \right) \tag{79}$$

The lower- and upper-bounds to maximize the sum-throughput are depicted in Fig. 9.

Fig. 10 and Fig. 11 are the counterparts of Figs. 8 and 9, respectively, for when $N_{\rm d} = K$.

To see more clearly the convergence of SINR with increasing M, in Fig. 12 the median of lower- and upper-bounds SINRs for $N_{\rm p} \to \infty$ (dashed lines) and $N_{\rm p} = 10K$ (solid lines) are depicted as a function of M. As you can see, the slopes of the curves are decreasing.

Appendix C

The sum spectral efficiency objective function in (59) is non-convex which leads us to the difference-of-convex optimizations method, which is locally optimal. In this appendix, we present a lower-bound that is a convex optimization problem and is a good estimate of the original problem for the

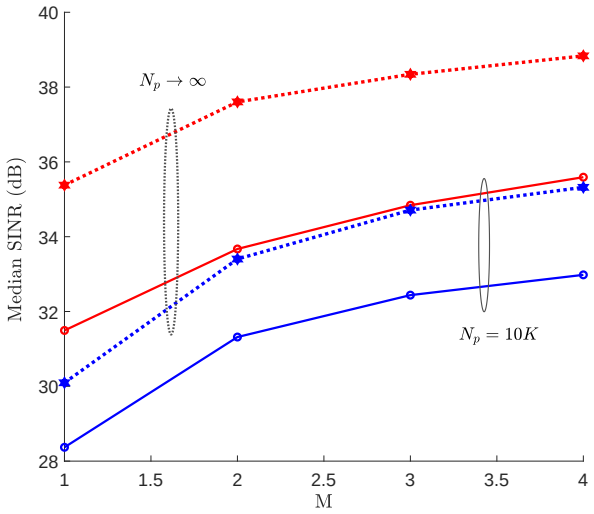

Fig. 12. Dashed lines, median of upper- and lower-bound SNRs for $N_{\rm p} \to \infty$ and different values of $M = 1, 2, 3, 4$. Solid lines, same quantities for $N_{\rm p} = 10K$

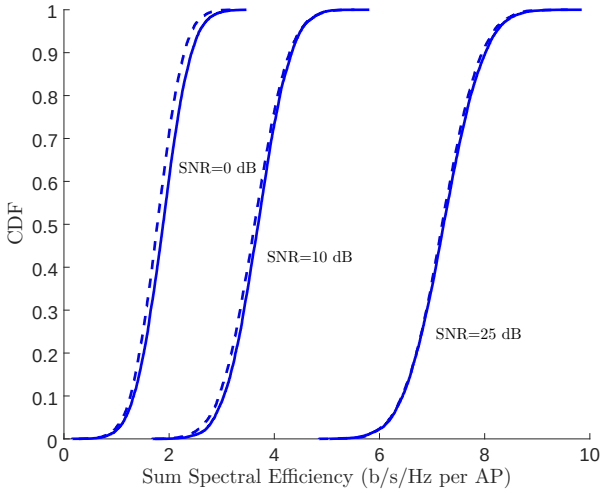

Fig. 13. CDF of the maximum gross sum spectral efficiency for $N_{\rm p} = 10K$ and three values of $\mathsf{SNR} = 0, 10, 25$ dB. For each case, the dashed curve corresponds to the convex lower-bound introduced in (82) and the solid curve is related to the maximum achievable sum spectral efficiency.

large SNRs. Consider the main optimization problem stated below

$$\max_{p_1,\ldots,p_K} \quad \sum_{k=1}^{K} \log_2\left(1 + \frac{P_{\rm t}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}\right) \quad \text{s.t.} \quad \boldsymbol{B}\boldsymbol{p} \leq \mathbf{1}. \tag{80}$$

Firstly, using AM-GM inequality:

$$1 + \frac{P_{\rm t}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell} \geq 2\sqrt{\frac{P_{\rm t}}{\sigma^2} \frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}}. \tag{81}$$

This leads to the following relaxed problem up to a multiplicative/additive constant.

$$\max_{p_1,\ldots,p_K} \quad \sum_{k=1}^{K} \log_2\left(\frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell}\right) \quad \text{s.t.} \quad \boldsymbol{B}\boldsymbol{p} \leq \mathbf{1}. \tag{82}$$

By introducing variables $\theta_k > 0,\ k = 1, 2, \ldots, K$, such that:

$$\frac{p_k}{1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell} \geq \frac{1}{\theta_k}, \tag{83}$$

we have:

$$p_k^{-1}\theta_k^{-1}\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right) \leq 1 \tag{84}$$

Therefore, by removing logarithm function as it is monotonic, the lower-bound problem in (82) is equivalent to

$$\min_{\substack{p_1,\ldots,p_K \\ \theta_1,\ldots,\theta_K}} \quad \prod_k \theta_k \quad \text{s.t.} \quad \boldsymbol{B}\boldsymbol{p} \leq \mathbf{1}, \quad p_k^{-1}\theta_k^{-1}\left(1 + \sum_{\ell=1}^{K} \gamma_{k,\ell}\, p_\ell\right) \leq 1,\ \forall k. \tag{85}$$

Since p_k's, θ_k's and $\gamma_{k,\ell}$'s are all non-negative, both cost function and constraints are posynomials, so the optimization problem (85) is a geometric optimization problem that can be turned into a convex problem and can even be solved directly with optimization packages like CVX.

In Fig. 13, the gaps between the solutions to (80) and (85) in the form of CDFs taken over the large-scale gains. Given $N = 20$, $K = 16$, and $N_{\rm p} = 10K$, three values of $\mathsf{SNR} = 0, 10, 25$ dB are entertained. It can be observed that obtained p_k's from (85) lead to the performance very close to the optimal value for different SNR values, although, as expected, lower-bound accuracy is better for larger SNR.

REFERENCES

[1] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. A. Valenzuela, "Is the PHY layer dead?" *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[2] X. Zhang and J. G. Andrews, "Downlink cellular network analysis with multi-slope path loss models," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1881–1894, May 2015.

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, 2014.

[4] S. Venkatesan, A. Lozano, and R. Valenzuela, "Network MIMO: Overcoming intercell interference in indoor wireless systems," in *Asilomar Conf. on Signals, Systems and Computers*, 2007, pp. 83–87.

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

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

[7] M. Attarifar, A. Abbasfar, and A. Lozano, "Random vs structured pilot assignment in cell-free massive MIMO wireless networks," in *IEEE Int'l Conf. Commun. Workshops (ICCW'18)*, 2018, pp. 1–6.

[8] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, 2018.

[9] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, pp. 616–619, 2019.

[10] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, 2020.

[11] N. Ross and D. Schuhmacher, "Wireless network signals with moderately correlated shadowing still appear Poisson," *IEEE Trans. Inform. Theory*, vol. 63, no. 2, pp. 1177–1198, 2016.

[12] B. Błaszczyszyn, M. K. Karray, and H. P. Keeler, "Wireless networks appear Poissonian due to strong shadowing," *IEEE Trans. on Wireless Communications*, vol. 14, no. 8, pp. 4379–4390, Aug. 2015.

[13] G. George, R. K. Mungara, A. Lozano, and M. Haenggi, "Ergodic spectral efficiency in MIMO cellular networks," *IEEE Trans. on Wireless Communications*, vol. 16, no. 5, pp. 2835–2849, 2017.

[14] 3GPP TS 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," 3GPP, Tech. Rep., Mar. 2017.

[15] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO communication*. Cambridge University Press, 2019.

[16] C. R. Rao and S. K. Mitra, "Generalized inverse of matrices and its applications," *John Wiley & Sons, New York*, 1971.

[17] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses." *IEEE Trans. Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

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

[19] M. Attarifar, A. Abbasfar, and A. Lozano, "Subset MMSE receivers for cell-free networks," *IEEE Trans. Wireless Commun.*, vol. 19, 2020.

[20] J. Ding and L. Huang, "On the continuity of generalized inverses of linear operators in hilbert spaces," *Linear algebra and its applications*, vol. 262, pp. 229–242, 1997.

[21] O. Y. Bursalioglu, C. Wang, H. Papadopoulos, and G. Caire, "RRH based massive MIMO with 'on the fly' pilot contamination control," in *IEEE Int'l Conf. on Communications (ICC'16)*, 2016, pp. 1–7.

[22] Y. Zhang, H. Cao, P. Zhong, C. Qi, and L. Yang, "Location-based greedy pilot assignment for cell-free massive MIMO systems," in *IEEE Int'l Conf. on Computer and Commun. (ICCC'18)*, Dec. 2018, pp. 392–396.

[23] R. Horst and N. V. Thoai, "Dc programming: overview," *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.

[24] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, 2016.