NOMA-Aided Cell-Free Massive MIMO with Underlay Spectrum-Sharing

Diluka Loku Galappaththige and Gayan Amarasuriya
Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA 62901
Email: {diluka.lg,gayan.baduge}@siu.edu

***Abstract*—We investigate the feasibility of employing non-orthogonal multiple-access (NOMA) in cell-free massive multiple-input multiple-output (MIMO) operating with underlay spectrum-sharing. In our proposed system model, multiple clusters of NOMA-enabled secondary users (SUs) are concurrently served by geographically distributed secondary access-points (S-APs) via conjugate beamforming. The uplink channels are estimated locally at each S-AP via pilots sent by SUs. A set of orthogonal pilots is shared among the secondary and primary clusters to strike a balance between the throughput and training overhead, while enabling massive connectivity in primary and secondary systems. We derive the achievable rates of the secondary system by capturing the adverse effects of inter/intra-cluster interference, primary/secondary pilot contamination, imperfect successive interference cancellation (SIC) and partial channel state information (CSI). We propose a transmit power allocation policy for the secondary system to mitigate the detrimental impact of near-far effects by virtue of max-min fairness criterion. Through an achievable rate analysis, we reveal that although the number of concurrently served users can be substantially boosted by employing the proposed system model, the achievable rates are adversely affected due to detection uncertainties with imperfect SIC and statistical CSI at NOMA-enabled SUs.**

I. Introduction

Cell-free massive multiple-input multiple-output (MIMO) can be used to render uniform service to many geographically distributed users [1], [2]. By virtue of inherent macro-diversity gains, reduced average propagation losses and improved spatial resolutions, cell-free massive MIMO has been shown to outperform the conventional co-located massive MIMO in terms of coverage/outage probability [1], [2]. Power-domain non-orthogonal multiple-access (NOMA) can be used to concurrently serve many users in the same time-frequency-spatial resource by adopting superposition-coded transmission and successive interference cancellation (SIC) decoding [3]–[5]. Moreover, a secondary system can be underlaid by exploiting underlay spectrum-sharing concepts to mitigate spectrum scarcity and underutilization/holes [6].

The coexistence of co-located massive MIMO and underlay spectrum-sharing is investigated in [7], while [8] extends orthogonal multiple-access (OMA) of [7] by adopting NOMA into the secondary system. In [9], the performance of OMA-based underlay spectrum-sharing aided cell-free massive MIMO is explored. In [10], the feasibility of exploiting NOMA for boosting massive connectivity of cell-free is studied. When cell-free massive MIMO is coupled with NOMA and underlay spectrum-sharing, not only spectral efficiency but also the number of simultaneous connections can be boosted. However, the coexistence of NOMA-aided cell-free massive MIMO with underlay spectrum-sharing has not yet been investigated. This present work is motivated by the aforementioned facts and gaps in NOMA literature.

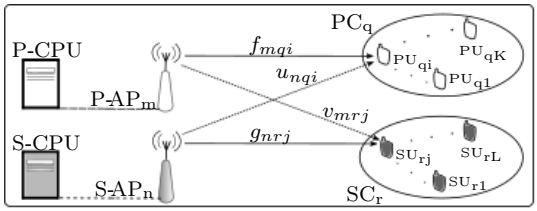

Fig. 1. Cell-free massive MIMO NOMA with underlay spectrum-sharing.

Our contribution: This paper serves as an extension to our previous work on OMA-based cell-free massive MIMO with underlay spectrum-sharing [9]. In this paper, we derive the achievable rates of NOMA-aided cell-free massive MIMO with underlay spectrum-sharing. A set of single-antenna primary access-points (P-APs) concurrently serves multiple clusters of primary users (PUs) via power-domain NOMA, while a secondary cell-free massive MIMO system is underlaid by operating with the same primary licensed spectrum. The uplink channel state information (CSI) is estimated locally at the secondary access-points (S-APs) by employing pilots sent by secondary users (SUs). By adopting conjugate precoding, S-APs simultaneously serve many SUs within multiple secondary clusters by using NOMA. The secondary interference inflicted at the PUs is controlled by employing a primary interference temperature at each S-AP. Deleterious near-far effect of the secondary system is mitigated by adopting a max-min fairness-based transmit power control policy. We analytically quantify the adverse effects of practical impediments such as primary/secondary pilot contamination, intra/inter-cluster interference, precoding uncertainty with contaminated CSI at S-APs, imperfect SIC decoding with partial/statistical CSI at SUs. Through our numerical results, we reveal that the number of concurrent connections can be significantly boosted by adopting the proposed system model. However, the achievable rates are detrimentally affected owing to the above practical transmission impairments. By virtue of the proposed max-min transmit power control, we show that the resulting common downlink rates at SUs ensure user-fairness by serving all users with a uniformly better quality-of-service than that of the equal power allocation.

II. System, channel and signal models

A. System and channel model

We consider a NOMA-aided cell-free massive MIMO system with underlay spectrum sharing (see Fig. 1) operating in time-division duplexing (TDD) mode. The primary system

consists of M single-antenna P-APs, which serve QK single-antenna PUs distributed in Q primary clusters by using its licensed frequency spectrum. Each primary cluster consists of K PUs. To boost overall spectrum efficiency by eliminating spectrum holes, a secondary system is underlaid within the same primary licensed spectrum. This secondary system comprises of N single-antenna S-APs serving RL SUs in R secondary clusters. These PUs and SUs are assigned to Q primary clusters and R secondary clusters based on their spatial directions [10], respectively. To ensure performance of the primary system is not hindered by the secondary CCI, the transmit power is constrained at each S-AP. Consequently, the secondary CCI inflected at the PUs always falls bellow a preset primary interference threshold. All P-APs and S-APs are connected to their respective primary and secondary central processing units (P-CPU/S-CPU) (see Fig. 1). It is assumed that the primary and secondary systems are operated in a synchronized manner, which is typical in underlay spectrum-sharing systems [2].

The channels from the mth P-AP to the kth PU in the qth primary cluster and the nth S-AP to the lth SU in the rth secondary cluster are denoted by f_{mqk} and g_{nrl}, respectively, where $m \in \{1, \cdots, M\}$, $q \in \{1, \cdots, Q\}$, $k \in \{1, \cdots, K\}$, $n \in \{1, \cdots, N\}$, $r \in \{1, \cdots, R\}$ and $l \in \{1, \cdots, L\}$. We represent the interference channel between the mth P-AP and the lth SU in the rth secondary cluster by v_{mrl}, while u_{nqk} denotes the channel between the nth S-AP and the kth PU in the qth primary cluster. These four channels can be modeled in a unified form as

$$h_{abc} = \zeta_{h_{abc}}^{1/2} \tilde{h}_{abc}, \tag{1}$$

where $h \in \{f, g, u, v\}$, $a \in \{m, n\}$, $b \in \{q, r\}$, and $c \in \{k, l\}$. Here, $\tilde{h}_{abc} \sim \mathcal{CN}(0, 1)$ captures quasi-static Rayleigh fading, and $\zeta_{h_{abc}}$ accounts for path-loss/shadowing, which is typically assumed to be fixed for many coherence intervals [2]. Hence, the channel statistics at the both P-APs and S-APs are assumed to be known a-prior because they change very slowly, and they need to be estimated once every tens/hundreds of coherence intervals [2].

B. Uplink channel estimation

As per the principles of cell-free massive MIMO [2], the uplink channels are estimated locally at each AP via user pilots [2]. To this end, τ_p symbols out of τ_c, which is the length of channel coherence interval, are allocated for pilot transmission. In pilot phase, all PUs/SUs concurrently transmit pilots to the P-APs/S-APs, and thereby, the respective channels can be estimated locally at each P-AP and S-AP. Owing to this, there is no need of exchanging channel estimates among the P-APs and S-APs [2]. To minimize training overhead, NOMA-enabled PUs/SUs located within the same primary clusters/secondary clusters are assigned with the same pilot sequence. Since only a limited number of orthogonal pilots is available for a given coherence interval [2], these pilots are shared among the primary and secondary systems. Thus, $W \leq \min(Q, R)$ pilots are assumed to be shared among the primary and secondary NOMA clusters, and the corresponding pilot assignment is defined as

$$\mathbf{\Phi}_P = \left[\mathbf{\Phi}; \tilde{\mathbf{\Phi}}_P\right] \quad \text{and} \quad \mathbf{\Phi}_S = \left[\mathbf{\Phi}; \tilde{\mathbf{\Phi}}_S\right]. \tag{2}$$

In (2), the set of orthogonal pilots that is shared by W primary/secondary clusters is denoted by $\mathbf{\Phi} \in \mathbb{C}^{W \times \tau_p}$, having a length of τ_p symbol durations. The remaining $(Q - W)$ primary clusters and $(R - W)$ secondary clusters are assigned with pilot sequences $\tilde{\mathbf{\Phi}}_P \in \mathbb{C}^{(Q-W) \times \tau_p}$ and $\tilde{\mathbf{\Phi}}_S \in \mathbb{C}^{(R-W) \times \tau_p}$, respectively. Since initial pilot set is orthogonal, we have $\mathbf{\Phi}\tilde{\mathbf{\Phi}}_P = \mathbf{0}$, $\mathbf{\Phi}\tilde{\mathbf{\Phi}}_S = \mathbf{0}$, $\tilde{\mathbf{\Phi}}_P\tilde{\mathbf{\Phi}}_S = \mathbf{0}$, $\mathbf{\Phi}_P = [\boldsymbol{\phi}_{P_1}^T, \cdots, \boldsymbol{\phi}_{P_q}^T, \cdots, \boldsymbol{\phi}_{P_Q}^T]^T$ and $\mathbf{\Phi}_S = [\boldsymbol{\phi}_{S_1}^T, \cdots, \boldsymbol{\phi}_{S_r}^T, \cdots \boldsymbol{\phi}_{S_R}^T]^T$, where $\boldsymbol{\phi}_{P_q} \in \mathbb{C}^{1 \times \tau_p}$ and $\boldsymbol{\phi}_{S_r} \in \mathbb{C}^{1 \times \tau_p}$ are the pilot sequences sent by the PUs in the qth primary cluster and the SUs in the rth secondary cluster, respectively. Moreover, $\|\boldsymbol{\phi}_{P_q}\|^2 = 1$ and $\|\boldsymbol{\phi}_{S_r}\|^2 = 1$ for $q \in \{1, \cdots, Q\}$ and $r \in \{1, \cdots, R\}$. Next, the pilot vector received at the nth S-AP can be written as

$$\mathbf{y}_{S_n} = \sqrt{\xi_p}\left(\sum_{r=1}^{R}\sum_{l=1}^{L} g_{nrl}\boldsymbol{\phi}_{S_r} + \sum_{q=1}^{Q}\sum_{k=1}^{K} u_{nqk}\boldsymbol{\phi}_{P_q}\right) + \mathbf{n}_{S_n}, \tag{3}$$

where $\xi_p = \tau_p P$ and P is the average transmitted pilot power at each PU/SU. Moreover, $\mathbf{n}_{S_n}$ is the additive white Gaussian noise (AWGN) vector at the nth S-AP, having identically and independently distributed (i.i.d.) $\mathcal{CN}(0, 1)$ elements.

By projecting $\boldsymbol{\phi}_{S_r}^H$ into (3), the sufficient statistics to estimate g_{nrj}, where $j \in \{1, \cdots, L\}$, can be obtained as

$$y_{S_{nr}} = \boldsymbol{\phi}_{S_r}^H \mathbf{y}_{S_n} = \sqrt{\xi_p}\left(\sum_{j=1}^{L} g_{nrj} + \sum_{i=1}^{K} u_{nri}\right) + n_{S_n}, \tag{4}$$

where $n_{S_n} = \boldsymbol{\phi}_{S_r}^H \mathbf{n}_{S_n} \sim \mathcal{CN}(0, 1)$ because $\boldsymbol{\phi}_{S_r}^H$ is a unitary vector. Then, the minimum mean square error (MMSE) estimate of g_{nrl} can be derived as (see Appendix A-A)

$$\hat{g}_{nrl} = \frac{\mathbb{E}\left[y_{S_{nr}}^* g_{nrl}\right]}{\mathbb{E}\left[|y_{S_{nr}}|^2\right]} y_{S_{nr}} = c_{S_{nrl}} y_{S_{nr}}, \tag{5}$$

where $c_{S_{nrl}}$ is given by

$$c_{S_{nrl}} \triangleq \frac{\sqrt{\xi_p}\zeta_{g_{nrl}}}{\xi_p\left(\sum_{j=1}^{L}\zeta_{g_{nrj}} + \sum_{i=1}^{K}\zeta_{u_{nri}}\right) + 1}. \tag{6}$$

By following a technique similar to (5), the MMSE estimate of f_{mqk} can be derived as

$$\hat{f}_{mqk} = \frac{\mathbb{E}\left[y_{P_{mq}}^* f_{mqk}\right]}{\mathbb{E}\left[|y_{P_{mq}}|^2\right]} y_{P_{mq}} = c_{P_{mqk}} y_{P_{mq}}, \tag{7}$$

where $y_{P_{mq}}$ and $c_{P_{mqk}}$ can be obtained by replacing secondary subscripts $\{S, n, r, l, j\}$ with respective primary subscripts $\{P, m, q, k, i\}$.

Owing to channel reciprocity property of TDD, P-APs/S-APs adopt locally estimated $\hat{f}_{mqk}$ and $\hat{g}_{nrl}$ as downlink CSI to design their precoders [2]. Since $y_{P_{mq}}$ and $y_{S_{nr}}$ are Gaussian distributed, we have $\hat{f}_{mqk} \sim \mathcal{CN}(0, \theta_{f_{mqk}})$ and $\hat{g}_{nrl} \sim \mathcal{CN}(0, \theta_{g_{nrl}})$, where $\theta_{f_{mqk}}$ and $\theta_{g_{nrl}}$ are given as

$$\mathbb{E}\left[|\hat{f}_{mqk}|^2\right] = \theta_{f_{mqk}} = \sqrt{\xi_p} c_{P_{mqk}} \zeta_{f_{mqk}}, \tag{8a}$$

$$\mathbb{E}\left[|\hat{g}_{nrl}|^2\right] = \theta_{g_{nrl}} = \sqrt{\xi_p} c_{S_{nrl}} \zeta_{g_{nrl}}. \tag{8b}$$

Then, the channel estimation errors pertaining to f_{mqk} and g_{nrl} are given by $\epsilon_{f_{mqk}} = f_{mqk} - \hat{f}_{mqk} \sim \mathcal{CN}(0, \zeta_{f_{mqk}} - \theta_{f_{mqk}})$ and $\epsilon_{g_{nrl}} = g_{nrl} - \hat{g}_{nrl} \sim \mathcal{CN}(0, \zeta_{g_{nrl}} - \theta_{g_{nrl}})$, respectively.

$$w_{S_{rl}} = \underbrace{\sqrt{P_S}\sum_{n=1}^{N}\sum_{j=1}^{l-1}\eta_{S_{nrj}}^{1/2} g_{nrl}\hat{g}_{nrj}^{*}\vartheta_{S_{rj}}}_{\text{Intra-cluster interference after SIC}} + \underbrace{\sqrt{P_S}\sum_{n=1}^{N}\sum_{j=l+1}^{L}\eta_{S_{nrj}}\left(g_{nrl}\hat{g}_{nrj}^{*}\vartheta_{S_{rj}} - \mathbb{E}\left[g_{nrl}\hat{g}_{nrj}^{*}\right]\hat{\vartheta}_{S_{rj}}\right)}_{\text{Error propagation due to imperfect SIC}}$$
$$+ \underbrace{\sqrt{P_S}\sum_{n=1}^{N}\sum_{r'\neq r}^{R}\sum_{j=1}^{L}\eta_{S_{nr'j}}^{1/2} g_{nrl}\hat{g}_{nr'j}^{*}\vartheta_{S_{r'j}}}_{\text{Intra-system interference}} + \underbrace{\sqrt{P_P}\sum_{m=1}^{M}\sum_{q=1}^{Q}\sum_{i=1}^{K}\eta_{P_{mqi}}^{1/2} v_{mrl}\hat{f}_{mqi}^{*}\vartheta_{P_{qi}}}_{\text{Inter-system interference}} + \underbrace{n_{P_{rl}}}_{\text{AWGN}} \quad (19)$$

C. Secondary Transmit Power Constraints

According to underlay spectrum sharing concepts [6], we constrain the transmit power at each S-AP such that the secondary CCI inflected at the PUs falls bellow the corresponding primary interference temperature at each PU. We define the total transmit power at the nth S-AP as

$$P_{S_n} \triangleq P_S \sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}, \quad (9)$$

where $\eta_{S_{nrj}}$ is the transmit power allocation coefficient at the nth S-AP for the jth SU in the rth secondary cluster, and it satisfies $\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}} \leq 1$. Moreover, P_S is the maximum allowable transmit power at each S-AP. Thus, the secondary CCI received at the kth PU in the qth primary cluster can be written as

$$y_{qk} = \sum_{n=1}^{N} u_{nqk} x_{S_n} = \sqrt{P_S}\sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}^{1/2} u_{nqk}\hat{g}_{nrj}^{*}\vartheta_{S_{rj}}, \quad (10)$$

where x_{S_n} is the transmitted signal by the nth S-AP (14). Next, the total average secondary CCI power ($P_{I_{qk}}$) inflected at the kth PU in the qth primary cluster can be derived as

$$P_{I_{qk}} = \mathbb{E}\left[|y_{qk}|^2\right] = \underbrace{P_S\sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}\mathbb{E}\left[\left|u_{nqk}\hat{g}_{nrj}^{*}\right|^2\right]}_{Z_{qk}}, \quad (11)$$

where Z_{qk} is derived as (see Appendix A-B)

$$Z_{qk} = \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}\theta_{g_{nrj}}\zeta_{u_{nqk}} + \sum_{n=1}^{N}\sum_{j=1}^{L}\eta_{S_{nqj}}\theta_{f_{mqk}}^{2}\left(\frac{\zeta_{u_{nqk}}\zeta_{g_{nqj}}}{\zeta_{f_{mqk}}}\right)^2. \quad (12)$$

Finally, we constrain the transmit power at the nth S-AP as

$$P_{S_n} = \min\left(P_S, \frac{I_{T_{11}}}{Z_{11}}, \cdots, \frac{I_{T_{qk}}}{Z_{qk}}, \cdots, \frac{I_{T_{QK}}}{Z_{QK}}\right), \quad (13)$$

where $I_{T_{qk}}$ is the interference temperature of the kth PU in the qth primary cluster.

D. Signal model

The S-APs employ conjugate precoding by using the locally estimated channel estimate in (5) to transmit the signal towards SUs in secondary clusters. The signal transmitted by the nth S-AP can be written as[1]

$$x_{S_n} = \sqrt{P_S}\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}^{1/2}\hat{g}_{nrj}^{*}\vartheta_{S_{rj}}, \quad (14)$$

where $\vartheta_{S_{rj}}$ is the signal intended for the jth SU in the rth secondary cluster satisfying $\mathbb{E}\left[|\vartheta_{S_{rj}}|^2\right] = 1$, and $\eta_{S_{nrj}}$ is selected to satisfy the transmit power constraint in (13). Then, we can write the received signals at the lth SU in the rth secondary cluster as

$$r_{S_{rl}} = \sum_{n=1}^{N} g_{nrl}x_{S_n} + \sum_{m=1}^{M} v_{mrl}x_{P_m} + n_{S_{rl}}, \quad (15)$$

where $n_{S_{rl}} \sim \mathcal{CN}(0,1)$ is an AWGN at the lth SU in the rth secondary cluster. To adopt the power-domain NOMA, SUs in the rth secondary cluster are assumed to be ordered based on the effective statistical channel gains as [4]

$$\mathbb{E}\left[\left|\sum_{n=1}^{N}\hat{g}_{nr1}\right|^2\right] \geq \cdots \geq \mathbb{E}\left[\left|\sum_{n=1}^{N}\hat{g}_{nrL}\right|^2\right]. \quad (16)$$

As per power-domain NOMA, the higher power levels are allocated for the users with lower statistical channel gains as

$$P_{S_{r1}} \leq \cdots \leq P_{S_{rL}}, \quad (17)$$

where $P_{S_{rj}} = P_S\eta_{S_{nrj}}$. Thus, the lth SU in the rth secondary cluster is able to decode the signals intended for the jth SUs for $\forall j(\geq l)$, given that it can decode its own signal. Then, the lth SU in the rth secondary cluster can successfully cancel the intra-cluster interference from the $\forall j(\geq l)$th SUs before decoding its own signal [4]. Furthermore, the lth SU treats the signals from all the users $j < l$ as interferences. The received signal given in (15) can be rearrange after SIC as

$$r_{S_{rl}} = \underbrace{\sqrt{P_S}\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2} g_{nrl}\hat{g}_{nrl}^{*}\vartheta_{S_{rl}}}_{\text{Desired signal}} + w_{S_{rl}}, \quad (18)$$

where $w_{S_{rl}}$ is an effective noise at the lth SU in the rth secondary cluster (19). In (19), $\hat{\vartheta}_{S_{rj}}$ is an estimate of $\vartheta_{S_{rj}}$. By assuming that $\vartheta_{S_{rj}}$ is drawn from a Gaussian distribution, $\hat{\vartheta}_{S_{rj}}$ is modeled to be jointly Gaussian distributed with a normalized correlation coefficient $\rho_{S_{rj}}$ as [10]

$$\vartheta_{S_{rj}} = \rho_{S_{rj}}\hat{\vartheta}_{S_{rj}} + e_{S_{rj}}, \quad (20)$$

where $\hat{\vartheta}_{S_{rj}}$ and $e_{S_{rj}}$ are statistically independent, and they are distributed as $\hat{\vartheta}_{S_{rj}} \sim \mathcal{CN}(0,1)$, $e_{S_{rj}} \sim \mathcal{CN}(0, \sigma_{e_{S_{rj}}}^2/(1+\sigma_{e_{S_{rj}}}^2))$, and $\rho_{S_{rj}} = 1/\sqrt{1+\sigma_{e_{S_{rj}}}^2}$.

III. Achievable Rate analysis

The received signal at the lth SU in the rth secondary cluster given in (18) can be rearranged to decode the desired signal by using statistical CSI as

$$r_{S_{rl}} = \underbrace{\sqrt{P_S}\,\mathbb{E}\left[\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2} g_{nrl}\hat{g}_{nrl}^{*}\right]}_{\text{Desired signal}}\vartheta_{S_{rl}} + \underbrace{\sqrt{P_S}\left(\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2} g_{nrl}\hat{g}_{nrl}^{*} - \mathbb{E}\left[\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2} g_{nrl}\hat{g}_{nrl}^{*}\right]\right)}_{\text{Detection uncertinity}}\vartheta_{S_{rl}} + w_{S_{rl}}, \quad (21)$$

where $w_{S_{rl}}$ is given in (19). We can now derive the signal-to-interference-plus-noise ratio (SINR) at the lth SU in the rth secondary cluster via (21) as

[1]The P-APs also adopt conjugate precoders designed via locally estimated CSI (7). The signal model and rate analysis for the primary system can be obtained by replacing subscripts $\{S, n, r, l, j\}$ pertaining to the secondary system by counterparts $\{P, m, q, k, i\}$ of the primary system. Performance analysis of the primary system is hence omitted for the sake of brevity.

$$\gamma_{S_{rl}} = P_S \left(\sum_{n=1}^{N} \eta_{S_{nrl}}^{1/2} \theta_{g_{nrl}} \right)^2 \Bigg/ \left(P_S \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{j=1}^{L} \eta_{S_{nrj}} \theta_{g_{nrj}} \zeta_{g_{nrl}} + P_S \sum_{j=1}^{3} I'_{S_{rj}} + P_P \sum_{m=1}^{M} \sum_{q=1}^{Q} \sum_{i=1}^{K} \eta_{P_{mqi}} \theta_{f_{mqi}} \zeta_{v_{mrl}} + 1 \right) \quad (24)$$

$$\gamma_{S_{rl}} = \frac{P_S \left| \mathbb{E}\left[\sum_{n=1}^{N} \eta_{S_{nrl}}^{1/2} g_{nrl} \hat{g}_{nrl}^{*} \right] \right|^2}{P_S \mathbb{V}\mathrm{ar}\left[\sum_{n=1}^{N} \eta_{S_{nrl}}^{1/2} g_{nrl} \hat{g}_{nrl}^{*} \right] + P_S \sum_{j=1}^{4} I_{S_{rj}} + \mathbb{E}\left[|n_{S_{rl}}|^2 \right]}, \quad (22)$$

where $I_{S_{rj}}$ for $j \in \{1,2,3,4\}$ can be defined as

$$I_{S_{r1}} = \mathbb{E}\left[\left| \sum_{n=1}^{N} \sum_{j=1}^{l-1} \eta_{S_{nrj}}^{1/2} g_{nrl} \hat{g}_{nrj}^{*} \right|^2 \right], \quad (23a)$$

$$I_{S_{r2}} = \mathbb{E}\left[\left| \sum_{n=1}^{N} \sum_{j=l+1}^{L} \eta_{S_{nrj}}^{1/2} \left(g_{nrl} \hat{g}_{nrj}^{*} \vartheta_{S_{rj}} - \mathbb{E}\left[g_{nrl} \hat{g}_{nrj}^{*} \right] \hat{\vartheta}_{S_{rj}} \right) \right|^2 \right], \quad (23b)$$

$$I_{S_{r3}} = \mathbb{E}\left[\left| \sum_{n=1}^{N} \sum_{r' \neq r}^{R} \sum_{j=1}^{L} \eta_{S_{nr'j}}^{1/2} g_{nrl} \hat{g}_{nr'j}^{*} \right|^2 \right], \quad (23c)$$

$$I_{S_{r4}} = \frac{P_P}{P_S} \mathbb{E}\left[\left| \sum_{m=1}^{M} \sum_{q=1}^{Q} \sum_{i=1}^{K} \eta_{P_{mqi}}^{1/2} v_{mrl} \hat{f}_{mqi}^{*} \right|^2 \right]. \quad (23d)$$

We compute the SINR by evaluating the expectation and variance terms in (22) as shown in (24) (see Appendix B). In (24), $I'_{S_{ri}}$ for $i \in \{1,2,3\}$ is given by

$$I'_{S_{r1}} = \sum_{n=1}^{N} \sum_{j=1}^{l-1} \eta_{S_{nrj}} \theta_{g_{nrl}}^2 \left(\frac{\zeta_{g_{nrj}}}{\zeta_{g_{nrl}}} \right)^2, \quad (25a)$$

$$I'_{S_{r2}} = 2 \sum_{n=1}^{N} \sum_{j=l+1}^{L} (1 - \rho_{S_{rj}}) \eta_{S_{nrj}} \theta_{g_{nrl}}^2 \left(\frac{\zeta_{g_{nrj}}}{\zeta_{g_{nrl}}} \right)^2, \quad (25b)$$

$$I'_{S_{r3}} = \frac{P_P}{P_S} \sum_{m=1}^{M} \sum_{i=1}^{K} \eta_{P_{mri}} \theta_{g_{nrl}}^2 \left(\frac{\zeta_{v_{mrl}} \zeta_{f_{mri}}}{\zeta_{g_{nrl}}} \right)^2. \quad (25c)$$

Thereby, we obtain the achievable rate of the lth SU in the rth secondary cluster as

$$R_{S_{rl}} = \left(\frac{\tau_c - \tau_p}{\tau_c} \right) \log_2 \left(1 + \gamma_{S_{rl}} \right), \quad (26)$$

where $\gamma_{S_{rl}}$ is given in (24), and the pre-log factor captures the effective portion of coherence interval utilized for payload data transmission. Finally, the total achievable sum rate of the secondary system is computed as

$$R_S = \sum_{r=1}^{R} \sum_{l=1}^{L} R_{S_{rl}}. \quad (27)$$

IV. Transmit Power Control

In cell-free massive MIMO, the achievable rates are detrimentally affected by near-far effect owing to distributed deployment of APs and users. Thus, user fairness must be guaranteed in terms of achievable rate in order to provide a uniform quality-of-service to all users. To this end, max-min power control algorithms have been shown to be optimal in the sense of user fairness in presence of near-far effects [2], [11]. Having been motivated by this fact, by invoking the max-min fairness concept [11], the transmit power allocation coefficients for S-APs can be computed to maximize the minimum achievable rate among all SUs located in all secondary clusters. Thus, we formulate a max-min transmit power control problem as

$$\underset{\eta_{S_{nrl}} \forall n,r,l}{\text{maximize}} \quad R_{S_{rl}} \quad (28a)$$

$$\text{subject to} \quad C_1 : \sum_{r=1}^{R} \sum_{j=1}^{L} \eta_{S_{nrj}} \theta_{g_{nrj}} \leq 1, \quad (28b)$$

$$C_2 : P_{S_n} \leq I_{T_{qk}} / Z_{qk}, \quad \forall q,k, \quad (28c)$$

$$C_3 : \eta_{S_{nrl}} \geq 0, \quad (28d)$$

where $R_{S_{rl}}$ is given in (26), and C_1 is obtained by using the maximum allowable transmit power at nth S-AP as follows:

$$\mathbb{E}\left[|x_{S_n}|^2 \right] \leq P_S \Rightarrow \mathbb{E}\left[\left| \sqrt{P_S} \sum_{r=1}^{R} \sum_{j=1}^{L} \eta_{S_{nrj}}^{1/2} \hat{g}_{nrj}^{*} \right|^2 \right] \leq P_S$$

$$\sum_{r=1}^{R} \sum_{j=1}^{L} \eta_{S_{nrj}} \mathbb{E}\left[|\hat{g}_{nrj}^{*}|^2 \right] \leq 1 \Rightarrow \sum_{r=1}^{R} \sum_{j=1}^{L} \eta_{S_{nrj}} \theta_{g_{nrj}} \leq 1, \quad (29)$$

where $\theta_{g_{nrj}}$ is given in (8b), and C_2 is evaluated by invoking the secondary transmit power constraint in (13). Since the achievable rate in (26) is a non-decreasing function of its argument, $R_{S_{rl}}$ can be replaced by $\gamma_{S_{rl}}$ given in (24). Then, by introducing a common SINR (λ) for all SUs and by defining a slack variable $\beta_{S_{nrl}} \triangleq \eta_{S_{nrl}}^{1/2}$, we equivalently reformulate the optimization problem in (28) as

$$\underset{\beta_{S_{nrl}} \forall n,r,l}{\text{maximize}} \quad \lambda \quad (30a)$$

$$\text{subject to} \quad \gamma_{S_{rl}} \geq \lambda$$

$$C_1 : \sum_{r=1}^{R} \sum_{j=1}^{L} \beta_{S_{nrj}}^2 \theta_{g_{nrj}} \leq 1, \quad (30b)$$

$$C_2 : P_{S_n} \leq I_{T_{qk}} / Z_{qk}, \quad \forall q,k, \quad (30c)$$

$$C_3 : \beta_{S_{nrl}} \geq 0, \quad (30d)$$

Since the objective function in (30) is quasi-concave functions the underlaying optimization problem is also quasi-concave [2]. Thus, the optimal solution can be found by using the Bisection method shown in Algorithm 1.

V. Numerical Results

Our simulation parameters are as follows: $\tau_c = 196$, $\tau_p = \max(Q,R)$, and $\zeta_{h_{abc}} = (d_0/d_{abc})^{\nu} \times 10^{\varphi_{abc}/10}$, where d_{abc} is the transmission distance, d_0 is the reference distance, ν is the path-loss exponent, and $10^{\varphi_{abc}/10}$ captures the shadow fading with $\varphi_{abc} \sim \mathcal{N}(0,8)$. The P-APs/S-APs are uniformly distributed, while PUs/SUs are randomly distributed over an area $800 \times 800\,\mathrm{m}^2$.

In Fig.2, the achievable sum rates of the primary and secondary systems are plotted as a function of the transmit power per P-AP. The maximum allowable transmit power at each S-AP is kept at $P_P/2$. The secondary transmit power control coefficients are calculated via the proposed max-min power allocation algorithm in Section IV. For the comparison purposes, the sum rates of the secondary system with equal power allocation have been plotted. Fig. 2 clearly reveals that the sum rates of primary NOMA clusters increase monotonically with unconstrained P_P. However, the sum rates of secondary clusters gradually increase up to a maximum in the low P_P regime and then decrease with increasing P_P. The

Algorithm 1 Bisection Algorithm

Input: Path-losses between P-APs/S-APs and PUs/SUs, and maximum allowable transmit powers of the PUs/SUs.

Output: The power control coefficients $\eta_{S_{nrl}}$ for $n \in \{1,\cdots,N\}$, $r \in \{1,\cdots,R\}$, and $l \in \{1,\cdots,L\}$.

Initialization: Define an initial region for the objective function by choosing appropriate values for λ_{min} and λ_{max}. Choose a tolerance $\epsilon > 0$.

1: **while** $\lambda_{max} - \lambda_{min} > \epsilon$ **do**

2: Calculate, $\lambda = (\lambda_{max} + \lambda_{min})/2$.

3: Solve the convex feasibility problem, which can be given as

$$\|\boldsymbol{V}_{\gamma_{S_{nrl}}}\| \leq \frac{1}{\sqrt{\lambda}} \left(\sum_{n=1}^{N} \beta_{S_{nrl}} \theta_{g_{nrl}} \right), \quad (31)$$

which is subjected to C_1, C_2, and C_3 given in (30b), (30c), and (30d), respectively. Moreover, $\boldsymbol{V}_{\gamma_{S_{nrl}}} \triangleq \left[\mathbf{v}_{S_1}^T,\ \mathbf{v}_{S_2}^T,\ 2\mathbf{v}_{S_3}^T,\ \frac{P_P}{P_S}\mathbf{v}_{S_4}^T,\ \frac{P_P}{P_S}\mathbf{v}_{S_5}^T,\ \frac{1}{\sqrt{P_S}} \right]$, where

$$\mathbf{v}_{S_1} \triangleq \left[\beta_{S_{111}} \sqrt{\theta_{g_{111}} \zeta_{g_{11l}}}, \cdots, \beta_{S_{NRL}} \sqrt{\theta_{g_{NRL}} \zeta_{g_{NRl}}} \right]^T, \quad (32a)$$

$$\mathbf{v}_{S_2} \triangleq \left[\frac{\beta_{S_{1r1}} \theta_{g_{1rl}} \zeta_{g_{1r1}}}{\zeta_{g_{1rl}}}, \cdots, \frac{\beta_{S_{Nr(l-1)}} \theta_{g_{Nrl}} \zeta_{g_{Nr(l-1)}}}{\zeta_{g_{Nrl}}} \right]^T, \quad (32b)$$

$$\mathbf{v}_{S_3} \triangleq \left[\frac{(1-\rho_{S_{r(l+1)}})\beta_{S_{1r(l+1)}}}{\zeta_{g_{1rl}} \theta_{g_{1rl}}^{-1} \zeta_{g_{1r(l+1)}}^{-1}}, \cdots, \frac{(1-\rho_{S_{rL}})\beta_{S_{NrL}}}{\zeta_{g_{Nrl}} \theta_{g_{Nrl}}^{-1} \zeta_{g_{NrL}}^{-1}} \right]^T \quad (32c)$$

$$\mathbf{v}_{S_4} \triangleq \left[\frac{\eta_{P_{1r1}}^{1/2} \theta_{g_{nrl}} \zeta_{v_{1rl}}}{\zeta_{g_{nrl}} \zeta_{f_{1r1}}^{-1}}, \cdots, \frac{\eta_{P_{MrK}}^{1/2} \theta_{g_{nrl}} \zeta_{v_{Mrl}}}{\zeta_{g_{nrl}} \zeta_{f_{MrK}}^{-1}} \right]^T, \quad (32d)$$

$$\mathbf{v}_{S_5} \triangleq \left[\sqrt{\eta_{P_{111}} \theta_{f_{111}} \zeta_{v_{1rl}}}, \cdots, \sqrt{\eta_{P_{MQK}} \theta_{f_{MQK}} \zeta_{v_{Mrl}}} \right]^T. \quad (32e)$$

4: If the problem is feasible, then set $\lambda_{min} = \lambda_\gamma$, otherwise set $\lambda_{max} = \lambda_\gamma$.

5: Stop if $\lambda_{max} - \lambda_{min} < \epsilon$. Otherwise go to Step 2.

6: **end while**

7: **return** $\eta_{S_{nrl}} = \beta_{S_{nrl}}^2$ for $n \in \{1,\cdots,N\}$, $r \in \{1,\cdots,R\}$, and $l \in \{1,\cdots,L\}$.

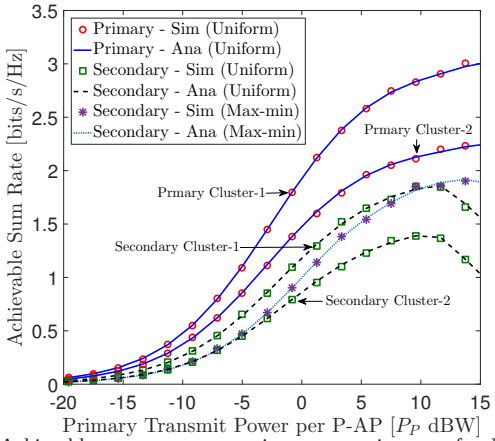

Fig. 2. Achievable sum rate versus primary transmit power for $M = N = 32$, $Q = R = 2$, $K = L = 2$, $I_{T_{qk}} = 0$ dB, $\sigma_P^2 = \sigma_S^2 = 1$, $\nu = 2.4$, $\rho_{P_{qi}} = \rho_{S_{rj}} = 0.2$, and $d_0 = 1$ m.

reason for this behavior is that when the secondary sum rates reach their maximum, the S-APs transmit at their maximum allowable power P_S. Thereafter, when P_P increases, the primary CCI at SUs keeps growing without bound. Hence, the secondary sum rates decrease as P_P grows larger. Fig. 2 depicts that the proposed max-min power allocation provides a common sum rate for all secondary clusters regardless of near-far effects and thus, guaranteeing user fairness in terms of achievable rates.

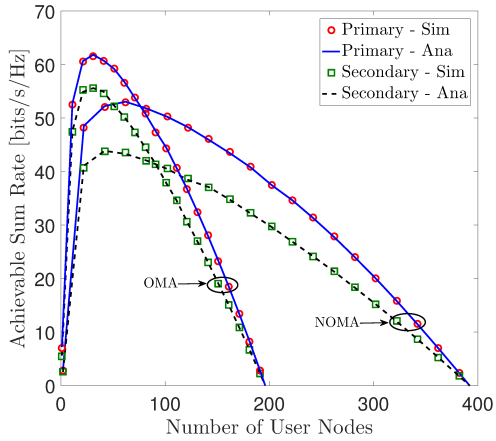

Fig. 3. Comparison between NOMA and OMA for $M = N = 32$, $P_P = 0$ dBW, $\sigma_P^2 = \sigma_S^2 = 1$, $\nu = 2.4$, $\rho_{P_{qi}} = \rho_{S_{rj}} = 0.2$, and $d_0 = 1$ m.

In Fig. 3, the achievable sum rates of the primary and secondary systems are plotted as a function of the number of concurrently served user nodes for both NOMA and OMA cases. Since the coherence interval $\tau_c = 196$, the maximum number of users that can be served via OMA is limited to 196. On the other hand, NOMA can serve more users than that of OMA because of each user cluster can consist of more than one user. However, Fig. 3 reveals that OMA outperforms NOMA in the regime of low number of users. The reason for this behavior is that the intra-cluster pilot contamination due to the shared pilots among NOMA clusters and the error propagation due to imperfect SIC, which hinder the performance of NOMA with respect to OMA.

VI. Conclusion

The feasibility of enabling massive connectivity by virtue of NOMA-aided cell-free massive MIMO with underlay spectrum-sharing has been investigated. The achievable rates of SUs and max-min power allocation coefficients at the S-APs have been quantified. We reveal that the number of concurrently served users can be significantly boosted by exploiting the coexistence of NOMA and underlay spectrum sharing. However, the residual interference yielded from pilot contamination, imperfect SIC, and partial CSI significantly hinders the achievable rates compared to a comparable system operating with conventional OMA. We show that the adverse near-far effect can be mitigated by virtue of serving all users with a common downlink rate via a max-min fairness-enabled transmit power allocation policy.

Appendix A

A. Derivation of the MMSE estimate

By substituting (4) in to (5), the MMSE estimate can be derived as [12]

$$\hat{g}_{nrl} = \frac{\mathbb{E}\left[\left(\sqrt{\xi_p} \left(\sum_{j=1}^{L} g_{nrj}^* + \sum_{i=1}^{K} u_{nri}^* \right) + n_{S_n}^* \right) g_{nrl} \right]}{\mathbb{E}\left[\left| \sqrt{\xi_p} \left(\sum_{j=1}^{L} g_{nrj} + \sum_{i=1}^{K} u_{nri} \right) + n_{S_n} \right|^2 \right]} y_{S_{nr}}$$

$$= \frac{\sqrt{\xi_p}\mathbb{E}\left[|g_{nrl}|^2 \right]}{\xi_p \left(\sum_{j=1}^{L} \mathbb{E}\left[|g_{nrj}|^2 \right] + \sum_{i=1}^{K} \mathbb{E}\left[|u_{nri}|^2 \right] \right) + \mathbb{E}\left[|n_{S_n}|^2 \right]} y_{S_{nr}}$$

$$= c_{S_{nrl}} y_{S_{nr}}, \quad (33)$$

where $c_{S_{nrl}}$ is defined in (6).

B. Derivation of Z_{qk} in (11)

Substituting (5) and (4) into (11), Z_{qk} can be computed as

$$\begin{aligned}Z_{qk} &= \sum_{n=1}^{N}\sum_{j=1}^{L}\eta_{S_{nqj}}c_{S_{nqj}}^2\mathbb{E}\left[\left|u_{nqk}y_{S_{nq}}^*\right|^2\right] \\ &\quad+\sum_{n=1}^{N}\sum_{r\neq q}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}c_{S_{nrj}}^2\mathbb{E}\left[\left|u_{nqk}y_{S_{nr}}^*\right|^2\right] \\ &= \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{j=1}^{L}\eta_{S_{nrj}}\theta_{g_{nrj}}\zeta_{u_{nqk}} \\ &\quad+\sum_{n=1}^{N}\sum_{j=1}^{L}\eta_{S_{nqj}}\theta_{f_{mqk}}^2\left(\frac{\zeta_{u_{nqk}}\zeta_{g_{nqj}}}{\zeta_{f_{mqk}}}\right)^2, \end{aligned} \tag{34}$$

where $\theta_{f_{mqi}}$ and $\theta_{g_{nrj}}$ are given in (8a) and (8b), respectively.

Appendix B
Derivation of SINR in (22)

Expectation term in the numerator of (22) is derived as

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2}g_{nrl}\hat{g}_{nrl}^*\right] &= \sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2}\mathbb{E}[(\hat{g}_{nrl}+\epsilon_{g_{nrl}})\,\hat{g}_{nrl}^*] \\ &= \sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2}\mathbb{E}\left[\left|\hat{g}_{nrl}\right|^2\right] = \sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2}\theta_{g_{nrl}}. \end{aligned} \tag{35}$$

Then, the variance term in (22) can be evaluated as

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\sum_{n=1}^{N}\eta_{S_{nrl}}^{1/2}g_{nrl}\hat{g}_{nrl}^*\right] \\ &= \sum_{n=1}^{N}\eta_{S_{nrl}}\left(\mathbb{E}\left[\left|(\hat{g}_{nrl}+\epsilon_{g_{nrl}})\,\hat{g}_{nrl}^*\right|^2\right]-\theta_{g_{nrl}}^2\right) \\ &= \sum_{n=1}^{N}\eta_{S_{nrl}}\left(2\theta_{g_{nrl}}^2+\theta_{g_{nrl}}(\zeta_{g_{nrl}}-\theta_{g_{nrl}})-\theta_{g_{nrl}}^2\right) \\ &= \sum_{n=1}^{N}\eta_{S_{nrl}}\theta_{g_{nrl}}\zeta_{g_{nrl}}. \end{aligned} \tag{36}$$

The expectation in (23a) can be derived as

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{n=1}^{N}\sum_{j=1}^{l-1}\eta_{S_{nrj}}^{1/2}g_{nrl}\hat{g}_{nrj}^*\right|^2\right] \\ &\overset{(a)}{=} \sum_{n=1}^{N}\sum_{j=1}^{l-1}\eta_{S_{nrj}}c_{S_{nrj}}^2\mathbb{E}\left[\left|g_{nrl}y_{S_{nr}}\right|^2\right] \\ &\overset{(b)}{=} \sum_{n=1}^{N}\sum_{j=1}^{l-1}\eta_{S_{nrj}}\theta_{g_{nrl}}\left(\theta_{g_{nrl}}+\zeta_{g_{nrl}}\right)\left(\frac{\zeta_{g_{nrj}}}{\zeta_{g_{nrl}}}\right)^2, \end{aligned} \tag{37}$$

where steps (a) and (b) are written by using (5) and (4), respectively, and then, evaluating expectations. The expectation term in (23b) can be computed as

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{n=1}^{N}\sum_{j=l+1}^{L}\eta_{S_{nrj}}^{1/2}\left(g_{nrl}\hat{g}_{nrj}^*\vartheta_{S_{rj}}-\mathbb{E}\left[g_{nrl}\hat{g}_{nrj}^*\right]\hat{\vartheta}_{S_{rj}}\right)\right|^2\right] \\ &= \sum_{n=1}^{N}\sum_{j=l+1}^{L}\eta_{S_{nrj}}\left(\mathbb{E}\left[\left|g_{nrl}\hat{g}_{nrj}^*\right|^2\right]+(1-2\rho_{S_{rj}})\left|\mathbb{E}[g_{nrl}\hat{g}_{nrj}]\right|^2\right) \\ &\overset{(c)}{=} \sum_{n=1}^{N}\sum_{j=l+1}^{L}\eta_{S_{nrj}}\theta_{g_{nrj}}\zeta_{g_{nrl}} \\ &\quad+2\sum_{n=1}^{N}\sum_{j=l+1}^{L}(1-\rho_{S_{rj}})\eta_{S_{nrj}}\theta_{g_{nrl}}^2\left(\frac{\zeta_{g_{nrj}}}{\zeta_{g_{nrl}}}\right)^2, \end{aligned} \tag{38}$$

where step (c) is written by using (5), (4), and (20). Then, the expectation term in (23c) can be derived by following steps similar to those used in (37) as

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{n=1}^{N}\sum_{r'\neq r}^{R}\sum_{j=1}^{L}\eta_{S_{nr'j}}^{1/2}g_{nrl}\hat{g}_{nr'j}^*\right|^2\right] \\ &= \sum_{n=1}^{N}\sum_{r'\neq r}^{R}\sum_{j=1}^{L}\eta_{S_{nr'j}}c_{S_{nr'j}}^2\mathbb{E}\left[\left|g_{nrl}y_{S_{nr'}}\right|^2\right] \\ &= \sum_{n=1}^{N}\sum_{r'\neq r}^{R}\sum_{j=1}^{L}\eta_{S_{nr'j}}\theta_{g_{nr'j}}\zeta_{g_{nrl}}. \end{aligned} \tag{39}$$

The final expectation in (23d) can be derived as follows:

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{m=1}^{M}\sum_{q=1}^{Q}\sum_{i=1}^{K}\eta_{P_{mqi}}^{1/2}v_{mrl}\hat{f}_{mqi}^*\right|^2\right] \\ &= \sum_{m=1}^{M}\sum_{i=1}^{K}\eta_{P_{mri}}c_{P_{mri}}^2\mathbb{E}\left[\left|v_{mrl}y_{P_{mr}}\right|^2\right] \\ &\quad+\sum_{m=1}^{M}\sum_{q\neq r}^{Q}\sum_{i=1}^{K}\eta_{P_{mqi}}\mathbb{E}\left[\left|v_{nrl}y_{P_{mq}}\right|^2\right] \\ &= \sum_{m=1}^{M}\sum_{q=1}^{Q}\sum_{i=1}^{K}\eta_{P_{mqi}}\theta_{f_{mqi}}\zeta_{v_{mrl}} \\ &\quad+\sum_{m=1}^{M}\sum_{i=1}^{K}\eta_{P_{mri}}\theta_{g_{nrl}}^2\left(\frac{\zeta_{v_{mrl}}\zeta_{f_{mri}}}{\zeta_{g_{nrl}}}\right)^2. \end{aligned} \tag{40}$$

By substituting (35), (36), (37), (38), (39), and (40) into (22), the desired SINR can be derived as given in (24).

References

[1] H. Q. Ngo *et al.*, "Cell-free Massive MIMO: Uniformly Great Service for Everyone," in *Proc. IEEE 16th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, June 2015, pp. 201–205.

[2] ——, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[3] Y. Saito *et al.*, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[4] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA with Limited Feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[5] Z. Ding *et al.*, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct 2017.

[6] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking Spectrum Gridlock with Cognitive Radios: An Information Theoretic Perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.

[7] H. Al-Hraishawi, G. A. Aruma Baduge, H. Q. Ngo, and E. G. Larsson, "Multi-Cell Massive MIMO Uplink with Underlay Spectrum Sharing," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 5, no. 1, pp. 119–137, Mar. 2019.

[8] Y. Li and G. A. Aruma Baduge, "Underlay Spectrum-Sharing Massive MIMO NOMA," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 116–119, Jan. 2019.

[9] D. L. Galappaththige and G. Amarasuriya, "Cell-Free Massive MIMO with Underlay Spectrum-Sharing," in *IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–7.

[10] Y. Li and G. A. Aruma Baduge, "NOMA-Aided Cell-Free Massive MIMO Systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec 2018.

[11] B. Radunovic and J. Le Boudec, "A Unified Framework for Max-Min and Min-Max Fairness with Applications," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 1073–1083, Oct 2007.

[12] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.