Complete Interference Mitigation Through Receiver-Caching in Wyner's Networks

Michèle Wigger[†], Roy Timo[‡], and Shlomo Shamai[†‡],
[†]LTCI CNRS, Telecom ParisTech, Université Paris-Saclay, 75013 Paris, France, michele.wigger@telecom-paristech.fr
[‡]Technische Universität München, roy.timo@tum.de, [†‡] Technion–Israel Institute of Technology, sshlomo@ee.technion.ac.il

***Abstract*—We present upper and lower bounds on the per-user multiplexing gain (MG) of Wyner's circular soft-handoff model and Wyner's circular full model with cognitive transmitters and receivers with cache memories. The bounds are tight for cache memories with prelog μ that exceeds $\frac{2}{3}D$ in the soft-handoff model and exceeds D in the full model, where D denotes the number of possibly demanded files. In these cases the per-user MG of the two models is $1+\frac{\mu}{D}$, the same as for non-interfering point-to-point links with caches at the receivers. Large receiver cache-memories thus allow to completely mitigate interference in these networks.**

I. Introduction

We consider a downlink communications problem in cellular networks, where each basestation is connected to a *cloud radio-access network (C-RAN)*. Such C-RANs can facilitate cooperative interference cancellation at the basestations, leading to higher data rates and improved energy efficiency [1]. With classical C-RAN architectures, however, interference cancellation requires that either the backhaul links connecting the basestations with the C-RAN have large capacity [2] or that the C-RAN can store every codebook and apply sophisticated signal processing techniques [3], [4]. Both approaches are not feasible in practical systems.

Researchers have also proposed to replace the C-RAN with a more complex *fog radio-access network (F-RAN)* [5], which can perform more sophisticated signal processing. It has been further proposed that the radio-head units in these F-RANs are equipped with memory spaces, so called *caches*, where they can pre-store information. Improvements in rates of such F-RAN architectures are presented, for example, in [6].

In this work, we consider the classical C-RAN architecture where the backhaul links are of limited capacities and the C-RAN cannot perform complex signal processing or store every codebook. We analyze the rates that are achievable when terminals have caches. In contrast to previous works on C-RANs or F-RANS that assigned caches to the basestations, e.g., [7], [8], [13], we assign caches to the receiving mobiles. (See also [9], [10], [11], [12] for results on noisy networks with receiver caching.) Notice that communicating cache contents from a central server to the mobiles is hardly more demanding than to the basestations when the communication happens during periods of weak network congestions, e.g., over night. This is possible when file popularities vary only slowly in time, which we will assume in this paper.

This work was supported by the Alexander von Humbold Foundation.

We model the communication from the base-stations to the mobiles by *Wyner's circular soft-handoff network* and *Wyner's circular full network*. We will show that when the receiving mobiles have sufficiently large cache memories, then in the high signal-to-noise ratio (SNR) regime a combination of

- coded caching [16],
- interference-cancelation at the receivers that is facilitated by the cache contents, and
- broadcast transmission at the transmitters,

allows one to completely mitigate interference. We present schemes that achieve the same *full* multiplexing gain (MG) as interference-free point-to-point links with caches at the receivers. For our scheme it suffices that each basestation is informed of only the messages intended to the two mobiles in its cell. The links connecting the C-RAN with the basestations can thus be of low rate, and there is no need for sophisticated signal processing or storing codebooks at the C-RAN.

Equipping mobile terminals in cellular networks with cache memories thus allows to substantially decrease backhaul costs and signal processing in C-RANs in the regime of high data rates when almost all interference needs to be mitigated.

II. Problem Description

We consider two models for cellular communication with K transmitter/receiver pairs:

Wyner's Circular Soft-Handoff Model (fig. 1): The signal sent at transmitter (Tx) k is observed at receiver (Rx) k and Rx $k+1$. Thus, if at time t, Tx k and Tx $k-1$ send the real-valued symbols $X_{k,t}$ and $X_{k-1,t}$, then Rx k observes

$$Y_{k,t} = X_{k,t} + \alpha_k X_{k-1,t} + Z_{k,t}, \tag{1}$$

where $\{Z_{k,t}\}$ is a sequence of independent and identically distributed (i.i.d.) real standard Gaussians; $\alpha_k \neq 0$ is a given real number; and $X_{0,t} = X_{K,t}$.

Wyner's Circular Full Model: In this model, the signal sent at Tx k is also observed at Rx $k-1$. Thus, at time t,

$$Y_{k,t} = X_{k,t} + \alpha X_{k-1,t} + \alpha X_{k+1,t} + Z_{k,t}, \tag{2}$$

where for simplicity we assume equal cross-gain $\alpha \neq 0$ for all receivers and $X_{K+1,t} = X_{1,t}$.

In both models, the inputs have to satisfy a symmetric average block-power constraint $P > 0$:

$$\frac{1}{n}\sum_{t=1}^{n} X_{k,t}^2 \leq P, \quad \forall k \in \{1,\dots,K\}, \quad \text{almost surely.} \tag{3}$$

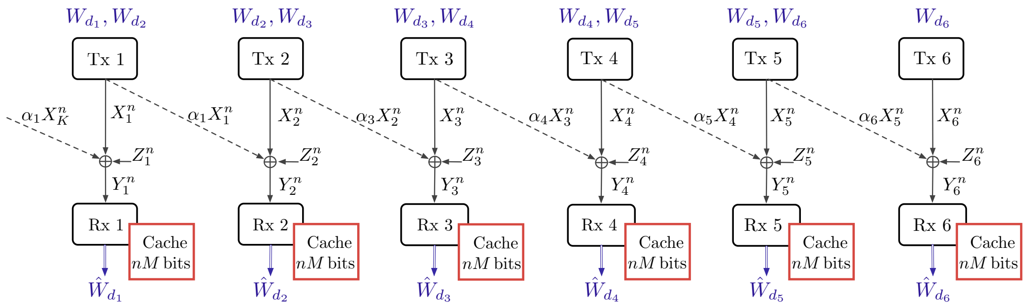


Fig. 1. Wyner's circular soft-handoff network for $K = 6$. Each receiver has a cache of nM bits. During the delivery phase each transmitter knows the messages demanded by the two receivers that observe its signal.

We consider a library of $D \geq 6$ messages:

$$\text{Library:} \qquad W_1, W_2, \dots, W_D, \tag{4}$$

which are independent and uniformly distributed over the set $\mathcal{W} \triangleq \{1, \dots, \lfloor 2^{nR} \rfloor\}$. Each receiver requests one of the files from the library. These requests, however, are not known prior to communications.

Finally, each receiver has a separate dedicated storage space, a *cache*, of size nM bits. A central server, who has access to the entire library (4), transmits dedicated caching information to each receiver, which this latter stores in its cache memory. This transmission takes place during an idle network period before the actual communication starts. As a consequence the transmission is error-free, but the transmitted information cannot depend on the receivers' actual demands, which are still unknown at this moment.

We describe the placement of the cache information and the general communication in more detail. There are three phases:

Phase 1 — Caching: A central server stores an arbitrary function of the entire library in this Rx k's cache, for $k \in \{1, \dots, K\}$. Let $\mathbb{V}_k$ denote the content in Rx k's cache,

$$\mathbb{V}_k = \phi_k(W_1, \dots, W_D), \tag{5}$$

for some $\phi_k \colon \mathcal{W}^D \to \{1, \dots, \lfloor 2^{nM} \rfloor\}$.

Phase 2 — Download from Server: Each Rx $k \in \{1, \dots, K\}$ communicates its demand $d_k \in \{1, \dots, D\}$ to its adjacent[1] transmitters, which then download the desired message W_{d_k} from the central server. So, after the server download phase, in the soft-handoff model:

$$\text{Tx } k \text{ knows:} \quad (W_{d_k},\ W_{d_{k+1}}), \tag{6}$$

and in the full model:

$$\text{Tx } k \text{ knows:} \quad (W_{d_{k-1}},\ W_{d_k},\ W_{d_{k+1}}), \tag{7}$$

Phase 3 — Delivery to Users: The transmitters communicate the demanded messages $W_{d_1}, W_{d_2}, \dots, W_{d_K}$ to Rx $1, \dots, K$, respectively. We assume that each transmitter and each receiver in the network knows the demand vector[2]

$$\mathbf{d} := (d_1, \dots, d_K). \tag{8}$$

[1]Tx j is adjacent to Rx k if the transmit signal of Tx j is observed at Rx k.
[2]Communicating $\mathbf{d} \in \{1, \dots, D\}^K$ to all terminals requires zero rate.

(Tx k however only knows messages W_{d_k} and $W_{d_{k+1}}$, (7).)

In the soft-handoff model, Tx k computes its inputs as

$$X_k^n = f_{\text{Soft},k}^{(n)}(W_{d_k}, W_{d_{k+1}}, \mathbf{d}), \tag{9}$$

and in the full model as

$$X_k^n = f_{\text{Full},k}^{(n)}(W_{d_{k-1}}, W_{d_k}, W_{d_{k+1}}, \mathbf{d}). \tag{10}$$

The encoding functions $f_{\text{Soft},k}^{(n)} \colon \mathcal{W}^2 \times \{1, \dots, D\}^K \to \mathbb{R}^n$ and $f_{\text{Full},k}^{(n)} \colon \mathcal{W}^3 \times \{1, \dots, D\}^K \to \mathbb{R}^n$ have to satisfy (3).

Each Rx k guesses its demanded message W_{d_k} as

$$\hat{W}_{d_k} = g_k^{(n)}(Y_k^n, \mathbb{V}_k, \mathbf{d}), \tag{11}$$

for some decoding function $g_k^{(n)} \colon \mathbb{R}^n \times \{1, \dots, \lfloor 2^{nM} \rfloor\} \times \{1, \dots, D\}^K \to \mathcal{W}$.

An error occurs in the communication whenever

$$\hat{W}_{d_k} \neq W_{d_k} \qquad \text{for any k} \in \{1, \dots, K\}. \tag{12}$$

Given power constraint P, a rate-memory pair $(R, \mathcal{M})$ is said *achievable*, if there exist encoding and decoding functions with vanishing probabilities of error as $n \to \infty$.

We perform a high-SNR analysis of our setup where $P \gg 1$. A per-user rate-memory MG pair $(\mathcal{S}, \mu)$ is said achievable, if for each power P there exists an achievable rate-memory pair (R_P, M_P) so that

$$\varlimsup_{K \to \infty} \varlimsup_{P \to \infty} \frac{1}{K} \cdot \frac{R_P}{\frac{1}{2}\log(1+P)} \geq \mathcal{S} \tag{13a}$$

$$\varlimsup_{P \to \infty} \cdot \frac{M_P}{\frac{1}{2}\log(1+P)} \geq \mu. \tag{13b}$$

We are interested in the *per-user rate-memory MG tradeoff*

$$\mathcal{S}^*(\mu) := \sup\Big\{\mathcal{S} \geq 0 \colon (\mathcal{S},\ \mu) \text{ an achievable MG pair}\Big\}.$$

We use subscripts $_{\text{Soft}}$ and $_{\text{Full}}$ to differentiate the per-user rate-memory MG tradeoffs of the two models.

III. Results

Let

$$\mathcal{S}_{\text{Soft,ach}}(\mu) := \begin{cases} \frac{2}{3} + \frac{3}{2}\frac{\mu}{D} & \text{if } 0 \leq \frac{\mu}{D} \leq \frac{2}{3} \\ 1 + \frac{\mu}{D} & \text{if } \frac{\mu}{D} \geq \frac{2}{3}. \end{cases} \tag{14}$$

Theorem 1 (Soft-Handoff Model)**.** *For Wyner's circular soft-handoff model:*

$$\min\left\{\frac{2}{3}+\frac{3\mu}{D}\ ,\ 1+\frac{\mu}{D}\right\}\geq \mathcal{S}^*_{\text{Soft}}(\mu)\geq \mathcal{S}_{\text{Soft,ach}}(\mu). \quad (15)$$

Proof: Converse omitted. Direct part in Section IV-A. ■

Corollary 1. *For the soft-handoff model and $\frac{\mu}{D}\geq\frac{2}{3}$:*

$$\mathcal{S}^*_{\text{Soft}}(\mu)=1+\frac{\mu}{D}. \quad (16)$$

Let

$$\mathcal{S}_{\text{Full,ach}}(\mu):=\begin{cases}\frac{2}{3}+\frac{4}{3}\frac{\mu}{D} & \text{if } 0\leq\frac{\mu}{D}\leq 1\\ 1+\frac{\mu}{D} & \text{if } \frac{\mu}{D}\geq 1.\end{cases} \quad (17)$$

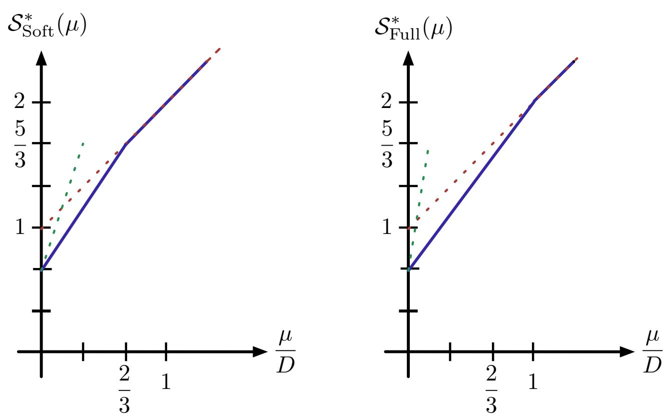
Theorem 2 (Full Model)**.** *For Wyner's circular full model:*

$$\min\left\{\frac{2}{3}+\frac{6\mu}{D}\ ,\ 1+\frac{\mu}{D}\right\}\geq \mathcal{S}^*_{\text{Full}}(\mu)\geq \mathcal{S}_{\text{Full,ach}}(\mu). \quad (18)$$

Proof: Converse omitted. Direct part in Section IV-B. ■

Corollary 2. *For the full model and for $\frac{\mu}{D}\geq 1$:*

$$\mathcal{S}^*_{\text{Full}}(\mu)=1+\frac{\mu}{D}. \quad (19)$$



(a) Soft-Handoff Model (b) Full Model

Fig. 2. Upper and lower bounds on the per-user rate-memory MG tradeoffs.

Figure 2 plots the upper and lower bounds on the per-user rate-memory MG tradeoff for the soft-handoff and the full circular Wyner model in Theorems 1 and 2. Without cache memories, $\mu=0$, the per-user MG is $\frac{2}{3}$ for both models.

In the regime $\frac{\mu}{D}\in[0,2/3]$, the achievable per-user rate-memory MG tradeoffs $\mathcal{S}_{\text{Soft,ach}}(\mu)$ and $\mathcal{S}_{\text{Full,ach}}(\mu)$ grow as a function of $\frac{\mu}{D}$ with slopes $\frac{3}{2}$ and $\frac{4}{3}$, respectively. The two slopes can be understood as follows. In the soft-handoff model, slope $\frac{3}{2}$ indicates that to achieve additional MG $\zeta\leq 1/3$ on the transmission over the network, *two* cache memories of size ζ are required: one to cancel the interference caused by this additional transmission, and one to cancel existing interference at the receiver that wants to decode the additional transmission. In the full model, *three* cache memories of size ζ are required to transmit an additional MG $\zeta\leq 1/3$ over the network: two cache memories are needed to cancel the symmetric interference caused by the additional transmission, and one to cancel existing interference at the receiver that wants to decode the additional transmission.

Finally, in both scenarios the per-user rate-memory MG tradeoff equals $1+\frac{\mu}{D}$ if the cache memories are sufficiently large (Corollaries 1 and 2). So, the same per-user rate-memory MG tradeoff is attained as in an interference-free network with receiver caching.

IV. Direct Parts

We have the following general proposition.

Proposition 1. *Consider an arbitrary interference network. If rate-memory pair (R,M) is achievable, then for every positive Δ the rate-memory pair*

$$\left(R+\frac{\Delta}{D},\ M+\Delta\right)\quad\text{is achievable.} \quad (20)$$

Proof. For each $d\in\{1,\ldots,D\}$, split message W_d into two submessages

$$W_d=(W_d^{(1)},W_d^{(2)}),$$

of rates R and $\frac{\Delta}{D}$. Apply the caching and delivery strategies of any scheme that achieves the rate-memory pair (R,M) to submessages $\{W_d^{(1)}\}_{d=1}^{D}$. Additionally, cache all submessages $W_1^{(2)},\ldots,W_D^{(2)}$ at each and every receiver. This requires an additional memory size of $n\Delta$ bits. □

A. Direct part for Theorem 1

The scheme in [2] achieves per-user rate-memory MG pair

$$\mathcal{S}_{\text{Soft}}=\frac{2}{3}\quad\text{and}\quad\mu=0. \quad (21)$$

In the following we show achievability of

$$\mathcal{S}_{\text{Soft}}=\frac{5}{3}\quad\text{and}\quad\mu=\frac{2}{3}D. \quad (22)$$

Time-sharing the schemes achieving (21) and (22) yields

$$\mathcal{S}^*_{\text{Soft}}(\mu)\geq\frac{2}{3}+\frac{3}{2}\cdot\frac{\mu}{D},\qquad\frac{\mu}{D}\in\left[0,\ \frac{2}{3}\right]. \quad (23)$$

Furthermore, by (22) and proposition 1,

$$\mathcal{S}^*(\mu)\geq 1+\frac{\mu}{D},\qquad\frac{\mu}{D}\in\left[\frac{2}{3},\infty\right), \quad (24)$$

which concludes the proof.

We now explain the scheme achieving (22). Define

$$\alpha_{\min}:=\min\{1,|\alpha_1|,\ldots,|\alpha_K|\}. \quad (25)$$

Fix a small $\epsilon>0$ and let the message rate R be

$$R=\frac{5}{3}\cdot\frac{1}{2}\log\left(1+\alpha^2_{\min}(P-\epsilon)\right)-5\epsilon. \quad (26)$$

Split each message into five independent submessages,

$$W_d=\left(W_d^{(1)},W_d^{(2)},W_d^{(3)},W_d^{(4)},W_d^{(5)}\right),\quad d\in\{1,\ldots,D\},$$

where each of these submessages is of equal rate

$$R/5=\frac{1}{3}\cdot\frac{1}{2}\log(1+\alpha^2_{\min}(P-\epsilon))-\epsilon.$$

We form a sixth *coded submessage*

$$W_d^{(6)} := W_d^{(1)} \bigoplus W_d^{(2)} \bigoplus W_d^{(3)} \bigoplus W_d^{(4)} \bigoplus W_d^{(5)},$$

where $\bigoplus$ denotes the x-or operation performed on the bit-string representation of the messages. In our scheme, each Rx k will recover arbitrary five submessages pertaining to its desired message W_{d_k}. It will then x-or these five submessages to obtain an estimate of the missing sixth submessage, which will allow it to form an estimate of its desired W_{d_k}.

Caching: The server makes the following cache assignment (see also figure 3) where $\mod$ denotes modulo operation:

$$\mathbb{V}_k = \begin{cases} \big(W_1^{(1)}, W_1^{(2)}, \dots, W_D^{(1)}, W_D^{(2)}\big), & \text{if } k \mod 3 = 1, \\ \big(W_1^{(3)}, W_1^{(4)}, \dots, W_D^{(3)}, W_D^{(4)}\big), & \text{if } k \mod 3 = 2, \\ \big(W_1^{(5)}, W_1^{(6)}, \dots, W_D^{(5)}, W_D^{(6)}\big), & \text{if } k \mod 3 = 0. \end{cases}$$

This cache assignment requires cache memory size

$$M = 2R/5 = \frac{2}{3} \cdot \frac{1}{2} \log(1 + \alpha_{\min}^2 (P - \epsilon)) - 2\epsilon. \tag{27}$$

Delivery to Users: Communication is split into periods 1–3 with $\lfloor \frac{n}{3} \rfloor$ consecutive channel uses each. In figures 3–5 we depict for each period the submessages sent by the transmitters and the submessages decoded at the receivers.

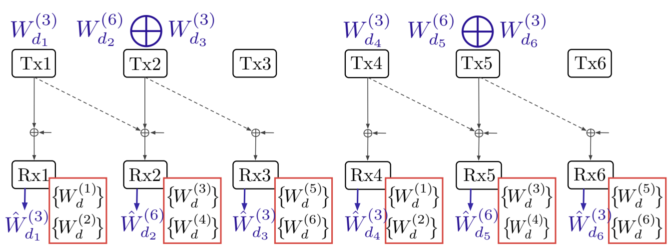


Fig. 3. Submessages sent during period 1.

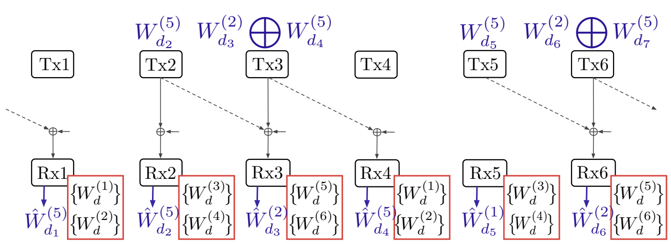


Fig. 4. Submessages sent during period 2.

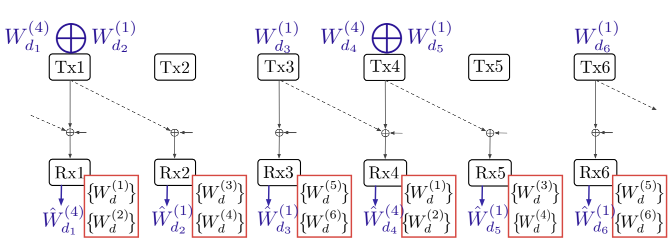


Fig. 5. Submessages sent during period 3.

In each period, every third transmitter sets its channel inputs to 0^n. In period 1, for example,

$$X_k^n = 0^n, \quad \text{if } k \mod 3 = 0.$$

Similarly, in period 2, $X_k^n = 0^n$ if $k \mod 3 = 1$ and in period 3, $X_k^n = 0^n$ if $k \mod 3 = 2$. In all three periods, also Tx K sets its channel inputs to 0^n. This decomposes the network into non-interfering subnets of two active transmitters and three receivers. (If $K \mod 3 = 2$, there is a last subnet with only Tx $K-1$ and Rx $K-1$.)

For example, in period 1, the first subnet comprises Txs 1–2 and Rxs 1–3, see figure 3. We now describe communication over this first subnet during period 1, assuming that d_1, d_2, d_3 are all different. (Otherwise, a slightly changed scheme achieves the same performance.)

Randomly draw for $k \in \{1, 2\}$ a Gaussian codebook $\mathcal{C}_k := \{X_k^{\lfloor n/3 \rfloor}(1), \dots, X_k^{\lfloor n/3 \rfloor}(\lfloor 2^{nR/5} \rfloor)\}$, of length $\lfloor \frac{n}{3} \rfloor$, rate-$3R/5$, and power $P - \epsilon$. Reveal the codebooks to all terminals of the subnet.

Tx 1 sends message $W_{d_1}^{(3)}$. To this end, it picks codeword

$$X_1^{\lfloor n/3 \rfloor}(W_{d_1}^{(3)}) \tag{28}$$

from its codebook $\mathcal{C}_1$ and sends it over the network.

Tx 2 forms $W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}$. It picks codeword

$$X_2^{\lfloor n/3 \rfloor}(W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}) \tag{29}$$

from its codebook $\mathcal{C}_2$ and sends it over the network.

Rx 1 optimally decodes submessage $W_{d_1}^{(3)}$ based on its outputs

$$Y_1^{\lfloor n/3 \rfloor} = X_1^{\lfloor n/3 \rfloor}(W_{d_1}^{(3)}) + Z_1^{\lfloor n/3 \rfloor}. \tag{30}$$

Rx 2 retrieves submessage $W_{d_1}^{(3)}$ from its cache and forms

$$\begin{aligned} \tilde{Y}_2^{\lfloor n/3 \rfloor} &:= Y_2^{\lfloor n/3 \rfloor} - \alpha_2 X_1^{\lfloor n/3 \rfloor}(W_{d_1}^{(3)}) \\ &= X_2^{\lfloor n/3 \rfloor}(W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}) + Z_1^{\lfloor n/3 \rfloor}. \end{aligned} \tag{31}$$

It optimally decodes the x-or submessage $W_{d_2}^{(1)} \bigoplus W_{d_3}^{(3)}$ based on $\tilde{Y}_2^{\lfloor n/3 \rfloor}$. Finally, it retrieves also $W_{d_3}^{(3)}$ from its cache and x-ors it with its guess of $W_{d_2}^{(1)} \bigoplus W_{d_3}^{(3)}$.

Recall that Tx 3 is deactivated. Rx 3 optimally decodes the x-or submessage $W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}$ based on its observations

$$Y_3^{\lfloor n/3 \rfloor} = \alpha_3 X_2^{\lfloor n/3 \rfloor}\left(W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}\right) + Z_3^{\lfloor n/3 \rfloor}. \tag{32}$$

It then retrieves $W_{d_2}^{(6)}$ from its cache and x-ors it with its guess of the x-or submessage $W_{d_2}^{(6)} \bigoplus W_{d_3}^{(3)}$.

The other subnets and other periods are treated analogously. This way, after periods 1–3, each Tx $k \in \{2, \dots, K-1\}$ has decoded three submessages of its desired W_{d_k}, see figures 3–5. It uses these three decoded submessages and the two submessages of W_{d_k} stored in its cache to reconstruct the missing sixth submessage and the original message W_{d_k}.

For example, Tx 2 has guessed $\hat{W}_{d_2}^{(6)}$ in period 1, $\hat{W}_{d_2}^{(5)}$ in period 2, and $\hat{W}_{d_2}^{(1)}$ in period 3 and has stored $W_{d_2}^{(3)}$ and $W_{d_2}^{(4)}$ in its cache memory. It forms

$$\hat{W}_{d_2}^{(2)} = \hat{W}_{d_2}^{(1)} \bigoplus W_{d_2}^{(3)} \bigoplus W_{d_2}^{(4)} \bigoplus \hat{W}_{d_2}^{(5)} \bigoplus \hat{W}_{d_2}^{(6)},$$

and produces as its final guess:

$$\hat{W}_{d_2} = \Big(\hat{W}_{d_2}^{(1)}, \hat{W}_{d_2}^{(2)}, W_{d_2}^{(3)}, W_{d_2}^{(4)}, \hat{W}_{d_2}^{(5)}\Big).$$

Analysis: Reconsider the first subnet of period 1. Each receiver 1–3 optimally decodes either submessage $W_{d_1}^{(3)}$ or the x-or submessage $\big(W_{d_2}^{(1)} \bigoplus W_{d_3}^{(3)}\big)$ based on interference-free Gaussian outputs (30)–(32). Since the rate of these submessages, $3R/5 = \frac{1}{2}\log(1+\alpha_{\min}^2 P) - 3\epsilon$, is below the capacity of the channels (30)–(32), the probability of decoding error tends to 0 as $n \to \infty$. If these decodings are error-free, then each Rx 1–3 produces the correct estimates in period 1.

Analogous conclusions hold for all subnets formed in periods 1–3. This proves that in the proposed scheme the probability of error at Rx 2 to Rx $K-1$ tends to 0 as $n \to \infty$.

If K is not a multiple of 3, the same does not apply to Rx 1 and Rx K because they have decoded less than three submessages. We can eliminate this edge effect by round-robin. Apply the same trick as before: split each message into $K-2$ message-parts and combine these $K-2$ message-parts into 2 new x-or parts so that from any $K-2$ parts one can reconstruct the original message. Now, time-share K instances of above scheme over K equally long super-periods, where in super-period $\ell \in \{1, \ldots, K\}$, we transmit the ℓ-th part of each message and we relabel Tx k and Rx k as Tx $k-\ell$ and Rx $k-\ell$, where indices need to be taken modulo K. This way each receiver is a bad receiver, i.e., Rx 1 or Rx K, only a fraction $\frac{K-2}{K}$ of the time, in which case it simply ignores its outputs.

The round-robin modification makes that the probability of decoding error now tends to 0 as $n \to \infty$ at all K receivers. But it requires that the rate of transmission is reduced by a factor $\frac{K-2}{K}$. This rate reduction however vanishes as $K \to \infty$.

B. Direct part for Theorem 2

The scheme in [2] achieves

$$\mathcal{S}_{\text{Full}} = \frac{2}{3} \quad \text{and} \quad \mu = 0. \tag{33}$$

We now describe a scheme that achieves

$$\mathcal{S}_{\text{Full}} = 2 \quad \text{and} \quad \mu = D. \tag{34}$$

Fix a small $\epsilon > 0$. Let

$$R = 2 \cdot \left(\frac{1}{2}\log(1 + P - \epsilon) - \epsilon\right), \tag{35}$$

and split each message W_d into two parts $\big(W_d^{(1)}, W_d^{(2)}\big)$ of equal rates $R/2$. Randomly draw for each $k \in \{1, \ldots, K\}$ a Gaussian codebook $\mathcal{C}_k := \big\{X_k^n(1), \ldots, X_k^n(\lfloor 2^{nR/2} \rfloor)\big\}$ of length n, rate $R/2$ and power $P - \epsilon$.
Caching: Cache messages $W_1^{(1)}, \ldots, W_D^{(1)}$ at all odd receivers and messages $W_1^{(2)}, \ldots, W_D^{(2)}$ at all even receivers.
Delivery to Users: For $k \in \{1, \ldots, K\}$ odd, Tx k sends the codeword $X_k^n\big(W_{d_k}^{(2)}\big)$ from codebook $\mathcal{C}_k$ over the channel. For $k \in \{1, \ldots, K\}$ even, Tx k sends the codeword $X_k^n\big(W_{d_k}^{(1)}\big)$ from codebook $\mathcal{C}_k$ over the channel.

For k odd, Rx k first takes messages $W_{d_{k-1}}^{(1)}$ and $W_{d_{k+1}}^{(1)}$ from its cache memory and forms

$$\begin{aligned} \hat{Y}_k^n &:= Y_k^n - \alpha X_{k-1}^n\big(W_{d_{k-1}}^{(1)}\big) - \alpha X_{k+1}^n\big(W_{d_{k+1}}^{(1)}\big) \\ &= X_k^n\big(W_{d_k}^{(2)}\big) + Z_k^n. \end{aligned} \tag{36}$$

It optimally decodes $W_{d_k}^{(2)}$ from these interference-free outputs $\hat{Y}_k^n$.

Similarly, for k even, Rx k first takes messages $W_{k-1}^{(2)}$ and $W_{k+1}^{(2)}$ from its cache memory and forms

$$\begin{aligned} \hat{Y}_k^n &:= Y_k^n - \alpha X_{k-1}^n\big(W_{d_{k-1}}^{(2)}\big) - \alpha X_{k+1}^n\big(W_{d_{k+1}}^{(2)}\big) \\ &= X_k^n\big(W_{d_k}^{(1)}\big) + Z_k^n. \end{aligned} \tag{37}$$

It optimally decodes $W_{d_k}^{(1)}$ from the new outputs $\hat{Y}_k^n$.

The probability of decoding error tends to 0 as $n \to \infty$ because the rate of communication $R/2$, (35), lies below the capacity of the created interference-free channels (36) and (37). Since our scheme requires cache memory $D \cdot \big(\frac{1}{2}\log(1 + P - \epsilon) - \epsilon\big)$, this proves achievability of (34).

Time-sharing the schemes achieving (33) and (34) yields

$$\mathcal{S}_{\text{Full}}^*(\mu) \geq \frac{2}{3} + \frac{4}{3} \cdot \frac{\mu}{D}, \qquad \frac{\mu}{D} \in \big[0,\ 1\big). \tag{38}$$

Finally, by (34) and proposition 1,

$$\mathcal{S}_{\text{Full}}^*(\mu) \geq 1 + \frac{\mu}{D}, \qquad \frac{\mu}{D} \in [1,\ \infty). \tag{39}$$

References

[1] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks." Online: `Arxiv: 1601.01070`

[2] A. Lapidoth, N. Levy, S. Shamai (Shitz), and M. Wigger, "Cognitive Wyner networks with clustered decoding," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6342-6367 Oct. 2014.

[3] M. Wigger, R. Timo, and S. Shamai, "Conferencing in Wyner's asymmetric interference network: effect of number of rounds," Online: `ArXiv:1603.05540`

[4] V. Ntranos, M. A. Maddah-Ali, and G. Caire, "Cellular interference alignment," *IEEE Trans. Inform. Theory,* vol. 61, no. 3, March 2015, pp. 1194–1217.

[5] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog computing based radio access networks: issues and challenges." Online: `ArXiv:1506.04233`

[6] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks." Online: `ArXiv:1601.02460`

[7] Y. Ugur, Z. H. Awan and A. Sezgin, "Cloud radio access networks with coded caching." Online: `ArXiv:1512.02385`.

[8] B. Azari, O. Simeone, U. Spagnolini and A. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching." To appear in *IEEE Wireless Comm. Letters*.

[9] S. Saeedi Bidokhti, R. Timo and M. Wigger, "Noisy Broadcast Networks with Receiver Caching." Online: `stanford.edu/ saeedi/jrnlcache.pdf`.

[10] P. Hassanzadeh, E. Erkip, J. Llorca,and A. Tulino, "Distortion-memory tradeoffs in Cache-aided wireless video delivery," in *Proc. of Allerton Conference* Monticello (IL), USA, Oct. 2015.

[11] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," in *Proc. of Allerton Conference,* Monticello (IL), USA, Oct. 2015.

[12] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback." Online: `ArXiv:1511.03961`.

[13] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE ISIT,* HongKong China, June 2015.

[14] S. Pooya Shariatpanah, S. Abolfazl Motahari, B. Hossein Khalaj, "Multi-server coded caching." Online: ArXiv:1503.00265.
[15] N. Naderializadeh, M. A. Maddah-Ali, A. S. Avestimehr, "Fundamental limits of cache-aided interference management." Online: ArXiv:1602.04207.
[16] M. A. Maddah-Ali, U. Niesen, "Fundamental limits of caching," in *IEEE Trans. on Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.