Gaussian Broadcast Channels with Receiver Cache Assignment

†Shirin Saeedi Bidokhti, ‡Michèle Wigger, and †‡Aylin Yener
† Stanford University, saeedi@stanford.edu, †‡The Pennsylvania State University & Stanford University, yener@engr.psu.edu
‡ LTCI, Telecom ParisTech, Université Paris-Saclay, 75013 Paris, France, michele.wigger@telecom-paristech.fr

***Abstract*—This paper considers a $K-$user Gaussian broadcast channel (BC) where receivers are equipped with cache memories. Lower and upper bounds are established on the *capacity-memory tradeoff*, i.e., the largest rate achievable for given cache-memories. The lower bound is based on a joint cache-channel coding scheme which generalizes the recently proposed piggyback coding to Gaussian BCs with unequal cache sizes. This paper also establishes lower and upper bounds on the *global capacity-memory tradeoff*, i.e., the maximum capacity-memory tradeoff over all possible cache assignments subject to a total cache memory constraint. The bounds match when the total cache memory is sufficiently large. It is shown that significantly larger rates can be achieved by carefully assigning larger cache memories to weaker receivers. In particular, cache allocation allows communication at rates that are (fundamentally) impossible to achieve with equal cache assignment. This shows the merit in carefully designing the cache size allocation in conjunction with channel qualities.**

I. Introduction

Introducing cache memory at various nodes in a network is a promising solution to increase performance of future communication networks. The idea is to prestore during periods of low network congestion popular content directly in cache memories at the end users or at closeby servers, so as to improve performance at peak-traffic periods. A main challenge is that when prestoring contents it is not yet known which files the users will request in the peak-traffic periods.

In traditional cache-aided system designs, the gain from caching is *local*; i.e., the gain is due to the fact that the transmitter does not need to send the fraction of the files already stored at the end users. Recently, Maddah-Ali and Niesen [1] showed that a smart design of the content of the caches lets the server send coded data in the delivery phase and simultaneously serve multiple users. This scheme thus offers a *global* caching gain that scales with the total memory sizes in the network, and is beyond the local caching gain. Following this work, improved caching and communication strategies have been proposed in [2]–[4], and fundamental converse results in [1, 5]–[8].

Maddah-Ali and Niesen model the peak-traffic communication channel by a noise-free rate-limited communication link. An alternative, and perhaps more realistic approach is to recognize that the channel is noisy and that the users have different channel statistics and observe noisy versions of the signal sent from the transmitter. In this work, we will take this approach and consider that the peak-traffic communication, i.e., the *delivery phase*, takes place over a Gaussian broadcast channel (BC). Noisy BCs with caching receivers have been addressed in other recent works, e.g., [9, 10, 12]–[16]. Under this model further global caching gains can be attained by *joint cache-channel coding* [9, 10] where the cache contents not only determine *what* messages the transmitter sends, but also *how* it transmits them. Joint cache-channel coding schemes were introduced in [9, 10] for erasure BCs, and have also been used in [11]. In this paper we extend the scheme in [9, 10], termed piggyback coding[1], to Gaussian BCs with unequal cache sizes. A main novelty of this work is thus the devise of a coding scheme for networks where receivers are equipped with unequal cache sizes. Most previous work has assumed either a uniform cache allocation across all users, or that some receivers have equal cache memory sizes and others have no caches at all [9, 10]. Exceptions are [17] and [18], but in these works the BC is a noise-free bitpipe (as in [1]).

The present paper also provides a new converse result under an arbitrary fixed cache assignment. It is often tighter than the previous converse results in [10, 12]. Moreover, it suggests that receivers can benefit from cache memories at weaker receivers but not at stronger receivers. This intuition is supported by our joint cache-channel coding schemes where thanks to the data stored in weak receivers cache memories, stronger receivers can piggyback information onto the communication to the weak receivers. These results thus suggest that assigning larger cache memories to weaker receivers is highly beneficial, more than in simply resolving the rate-bottleneck caused by the weak receivers.

To make this statement more precise, we derive upper and lower bounds on the *global capacity-memory tradeoff* of the Gaussian BC, where one is allowed to optimize over the cache assignment subject to a total cache constraint. The bounds are generally close and meet when the total cache memory exceeds a certain threshold. Numerical evaluation of the bounds suggest that this global capacity-memory tradeoff is substantially larger than the capacity-memory tradeoff under a uniform cache assignment.

II. Problem Definition

Consider a transmitter and K receivers $k = 1, \ldots, K$. The transmitter has access to a library with D independent files

[1]By recasting the cache contents as side-information, piggyback coding can be seen as a simplified version of the joint source-channel coding scheme for BCs in [20].

(messages) $W_1, \ldots, W_D$, each uniformly distributed over the set[2] $\{1, \ldots, 2^{nR}\}$. So, $R \geq 0$ denotes the rate of transmission and n the transmission blocklength.

Each receiver is equipped with a cache memory, and communication takes place in two phases: a *caching phase* (also called placement phase) that occurs before the receivers demands are known, and a *delivery phase* after receivers request messages of their interest. We denote the message demanded by receiver k with W_{d_k}, $d_k \in \{1, \ldots, D\}$, and refer to the vector $(d_1, \ldots, d_K)$ as the demand vector $\mathbf{d}$.

Each receiver $k \in \{1, \ldots, K\}$ is equipped with a cache of certain size, described by a nonnegative integer M_k. In the caching phase, the demand vector $\mathbf{d}$ is not known. The transmitter places in cache k

$$\mathbb{V}_k := g_k(W_1, \ldots, W_D), \tag{1}$$

for some function $g_k : \{1, \ldots, 2^{nR}\}^D \to \{1, \ldots, 2^{n\mathsf{M}_k}\}$.

The caching phase occurs during a low congestion period, and as in all caching models we too assume that $\mathbb{V}_k$ is reliably conveyed to receiver k's cache, for each $k \in \{1, \ldots, K\}$.

The delivery phase occurs during a high congestion period, which we model by a Gaussian BC. So in channel use t receiver k's output is

$$Y_{k,t} = X_t + Z_{k,t}, \tag{2}$$

where X_t is the input to the channel and $\{Z_{k,t}\}$ is a Gaussian random variable with zero mean and variance $\sigma_k^2 > 0$. The channel inputs are subject to an average block-power constraint P and we assume without loss of generality

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_K^2.$$

At the beginning of the delivery phase, each receiver k demands message W_{d_k}, $d_k \in \{1, \ldots, D\}$, and the transmitter and all the receivers get informed of the demand vector $\mathbf{d} = (d_1, \ldots, d_K)$. Using this information, the transmitter forms the channel input sequence $X^n = (X_1, \ldots, X_n)$ as

$$X^n := f_{\mathbf{d}}(W_1, \ldots, W_D) \tag{3}$$

for some encoding function $f_{\mathbf{d}} : \{1, \ldots, 2^{nR}\}^D \to \mathcal{X}^n$.

Receiver $k \in \{1, \ldots, K\}$ observes the channel output sequence $Y_k^n = (Y_{k,1}, \ldots, Y_{k,n})$. Given demand vector $\mathbf{d}$, cache content $\mathbb{V}_k$ and channel outputs Y_k^n, it produces

$$\hat{W}_k := \varphi_{k,\mathbf{d}}(Y_k^n, \mathbb{V}_k), \tag{4}$$

its estimate of the desired message W_{d_k}, by means of a decoding function $\varphi_{k,\mathbf{d}} : \mathcal{Y}_k^n \times \{1, \ldots, 2^{n\mathsf{M}_k}\} \to \{1, \ldots, 2^{nR}\}$.

The worst-case probability of error at any receiver and any demand $\mathbf{d}$ is given by

$$\mathsf{P}_\mathsf{e} := \mathbb{P}\Big[\bigcup_{\mathbf{d} \in \mathcal{D}} \bigcup_{k=1}^{K} \{\hat{W}_k \neq W_{d_k}\} \Big].$$

A rate-memory tuple $(R, \mathsf{M}_1, \ldots, \mathsf{M}_K)$ is said *achievable* if for any $\epsilon > 0$ there exists a sufficiently large blocklength n and caching, encoding, and decoding functions as in (1), (3), and (4) so that $\mathsf{P}_\mathsf{e} \leq \epsilon$.

Definition 1: The *capacity-memory tradeoff* $\mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K)$ is the largest symmetric rate R for which the rate-memory tuple $(R, \mathsf{M}_1, \ldots, \mathsf{M}_K)$ is achievable:

$$\mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K) := \sup\{R\colon (R, \mathsf{M}_1, \ldots, \mathsf{M}_K) \text{ achievable}\}.$$

The main goal is to optimize the cache assignment $(\mathsf{M}_1, \ldots, \mathsf{M}_K)$ to attain the largest capacity-memory tradeoff $\mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K)$ under the total cache constraint:

$$\sum_{k=1}^{K} \mathsf{M}_k \leq \mathsf{M}. \tag{5}$$

Definition 2: The *global capacity-memory tradeoff* $\mathsf{C}^\star(\mathsf{M})$ is defined as:

$$\mathsf{C}^\star(\mathsf{M}) := \max_{\substack{\mathsf{M}_1, \ldots, \mathsf{M}_K > 0: \\ \sum_{k=1}^K \mathsf{M}_k \leq \mathsf{M}}} \mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K). \tag{6}$$

III. Results

A. Preliminaries and Notation

In the absence of cache memories, $\mathsf{M}_1 = \ldots = \mathsf{M}_2 = 0$, the capacity-memory tradeoff $\mathsf{C}(\mathsf{M}_1 = 0, \ldots, \mathsf{M}_K = 0)$ is well known: It is the largest symmetric rate R with which K independent messages can be reliabily sent to the K users and is given as follows.

$$\begin{aligned} \mathsf{C}_0 &:= \mathsf{C}(\mathsf{M}_1 = 0, \ldots, \mathsf{M}_K = 0) \\ &= \frac{1}{2} \log_2 \left(1 + \frac{\beta_1 P}{\sum_{k=2}^K \beta_k P + \sigma_1^2} \right) \end{aligned} \tag{7}$$

for the unique choice of parameters $\beta_1, \ldots, \beta_K \geq 0$ summing to 1 and satisfying

$$\frac{\beta_1 P}{\sum_{k=2}^K \beta_k P + \sigma_1^2} = \frac{\beta_i P}{\sum_{k=i+1}^K \beta_k P + \sigma_i^2}, \quad \forall i \in \{2, \ldots, K\}.$$

We will also need the following notation. For a subset $\mathcal{S} \subseteq \{1, \ldots, K\}$ of receivers, $\mathsf{C}_\mathcal{S}$ denotes the largest symmetric rate that can be achieved in a BC with receivers in $\mathcal{S}$ assuming no cache memories. Let

$$\mathcal{S} := \{s_1, \ldots, s_{|\mathcal{S}|}\} \subseteq \{1, \ldots, K\}, \ s_1 < \cdots < s_{|\mathcal{S}|}. \tag{8}$$

We have

$$\mathsf{C}_\mathcal{S} := \frac{1}{2} \log_2 \left(1 + \frac{\beta_1 P}{\sum_{k=2}^{|\mathcal{S}|} \beta_k P + \sigma_1^2} \right), \tag{9}$$

where $\beta_1, \ldots, \beta_{|\mathcal{S}|}$ form the unique choice of $|\mathcal{S}|$ real numbers in $[0, 1]$ that sum to 1 and satisfy

$$\frac{\beta_1 P}{\sum_{k=2}^K \beta_k P + \sigma_1^2} = \frac{\beta_i P}{\sum_{k=i+1}^K \beta_k P + \sigma_{s_i}^2}, \qquad \forall i \in \mathcal{S}. \tag{10}$$

We will use the abbreviation

$$\mathsf{C}_k := \mathsf{C}_{\{k\}} = \frac{1}{2} \log \left(1 + \frac{P}{\sigma_k^2} \right), \quad k \in \{1, \ldots, K\}. \tag{11}$$

Notice $\mathsf{C}_{\{1,\ldots,K\}} = \mathsf{C}_0$.

[2]For simplicity, 2^{nR_d} is assumed to be integer.

Throughout the paper, we will use the following short-hand notations for any choice of $t \in \{1,\ldots,K-1\}$:

$$\tau^{(t)} := \binom{K}{t}, \tag{12}$$

$$N^{(t)} := \binom{K}{t+1}. \tag{13}$$

Let $\mathcal{G}_1^{(t)}, \mathcal{G}_2^{(t)}, \ldots, \mathcal{G}_{\tau^{(t)}}^{(t)}$ denote the $\tau^{(t)}$ subsets of $\{1,\ldots,K\}$ that are of size t, and $\mathcal{S}_1^{(t)}, \mathcal{S}_2^{(t)}, \ldots, \mathcal{S}_{\tau^{(t+1)}}^{(t)}$ the $N^{(t)}$ subsets of $\{1,\ldots,K\}$ that are of size $t+1$. We define

$$\mu^{(t)} := \sum_{j=1}^{N^{(t)}} \prod_{k \in \bar{\mathcal{S}}_j^{(t)}} \mathsf{C}_k, \tag{14}$$

where $\bar{\mathcal{S}}_j^{(t)} := \{1,\ldots,K\} \backslash \mathcal{S}_j^{(t)}$.

B. Lower Bounds on Capacity-Memory Tradeoffs

Similarly to [19], we have:

Proposition 1 (Local caching gain): For any $\Delta > 0$,

$$\mathsf{C}^\star(\mathsf{M}+\Delta) \geq \mathsf{C}^\star(\mathsf{M}) + \frac{\Delta}{K \cdot D}. \tag{15}$$

In Sections IV and V ahead we present two coding schemes for appropriate choices of cache memory allocations. They immediately yield two lower bounds on the global capacity-memory tradeoff, see the following Theorems 2 and 3.

Let $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$ be nonnegative parameters summing to 1 and so that the following $K-1$ inequalities hold:

$$\frac{1}{2}\log\left(1+\frac{(\tilde{\beta}_1+\tilde{\beta}_2)P}{\sigma_2^2+\sum_{i=3}^K \tilde{\beta}_i P}\right) - \frac{1}{2}\log\left(1+\frac{P}{\sigma_1^2}\right) = \frac{1}{2}\log\left(\frac{\tilde{\beta}_k P}{\sigma_k^2+\sum_{i=k+1}^K \tilde{\beta}_i P}\right), \quad \forall k \in \{2,\ldots,K\}. \tag{16}$$

Theorem 2: Given cache memory

$$\mathsf{M} := \frac{D \cdot \mathsf{C}_{\{2,\ldots,K\}}}{\mathsf{C}_{\{2,\ldots,K\}} + \mathsf{C}_1} \cdot \left(\frac{1}{2}\log\left(1+\frac{(\tilde{\beta}_1+\tilde{\beta}_2)P}{\sigma_2^2\sum_{i=3}^K \tilde{\beta}_i P}\right) - \mathsf{C}_1\right), \tag{17}$$

the global capacity-memory tradeoff is lower bounded as

$$\mathsf{C}(\mathsf{M}_1,\ldots,\mathsf{M}_K) \geq \frac{\mathsf{C}_{\{2,\ldots,K\}}\mathsf{C}_1}{\mathsf{C}_{\{2,\ldots,K\}}+\mathsf{C}_1} + \frac{M}{D}. \tag{18}$$

Proof: See the scheme in Section IV. ∎

Theorem 3: For each parameter $t \in \{1,\ldots,K-1\}$, given a total cache memory

$$\mathsf{M}^{(t)} = D\frac{t}{\mu^{(t)}} \sum_{\ell=1}^{\tau^{(t)}} \left(\prod_{k \in \bar{\mathcal{G}}_\ell^{(t)}} \mathsf{C}_k\right), \tag{19a}$$

the global capacity-memory tradeoff is lower bounded by

$$\mathsf{C}^\star\big(\mathsf{M}^{(t)}\big) \geq \frac{1}{\mu^{(t)}} \sum_{\ell=1}^{\tau^{(t)}} \left(\prod_{k \in \bar{\mathcal{G}}_\ell^{(t)}} \mathsf{C}_k\right), \tag{19b}$$

where $\bar{\mathcal{G}}_\ell^{(t)} := \{1,\ldots,K\} \backslash \mathcal{G}_\ell^{(t)}$.

Proof: See the scheme in Section V. ∎

C. Upper Bounds on the Global Capacity-Memory Trade-off

We present two upper bounds. The first is based on an upper bound on the capacity-memory tradeoff in [10, Theorem 9]. The second requires proving a new converse and is often tighter, but more cumbersome to evaluate, than the first bound.

Theorem 4: For each $t \in \{1,\ldots,K\}$, the global capacity-memory tradeoff is upper bounded by

$$\mathsf{C}^\star(\mathsf{M}) \leq \frac{1}{\tau^{(t)}} \sum_{\ell=1}^{\tau^{(t)}} \mathsf{C}_{0,\mathcal{G}_\ell^{(t)}} + \frac{t \cdot \mathsf{M}}{K \cdot D}. \tag{20}$$

Proof: Fix $t \in \{1,\ldots,K\}$. Specialize Theorem 9 in [10] to $\mathcal{S} = \mathcal{G}_\ell^{(t)}$, for $\ell = 1,\ldots\tau^{(t)}$, and take the average of the $\tau^{(t)}$ obtained constraints. ∎

Specializing Theorem 4 to $t=1$ and to $t=K$ results in the following corollary.

Corollary 4.1:

$$\mathsf{C}^\star(\mathsf{M}) \leq \mathsf{C}_0 + \frac{\mathsf{M}}{D} \tag{21a}$$

$$\mathsf{C}^\star(\mathsf{M}) \leq \sum_{k=1}^K \frac{\mathsf{C}_k}{K} + \frac{\mathsf{M}}{K \cdot D}. \tag{21b}$$

The upper bound in (21a) performs better for small values of M and the upper bound in (21b) is better for large values of M. The latter is in fact tight when M exceeds a certain threshold, see Corollary 6.1 ahead.

Before presenting our second upper bound, we introduce some notation. Consider fixed cache memory sizes $\mathsf{M}_1,\ldots,\mathsf{M}_K$. For any subset of receivers $\mathcal{S}$ as defined in (8), let $R_{\mathcal{S}}^\star$ be the largest rate so that the rate tuple

$$\left(R-\alpha_{\mathcal{S},1}, R-\alpha_{\mathcal{S},2}, \ldots, R-\alpha_{\mathcal{S},|\mathcal{S}|}\right),$$

where

$$\alpha_{\mathcal{S},k} = \frac{\sum_{i=1}^k \mathsf{M}_{s_i}}{D-k+1}, \tag{22}$$

lies in the capacity region of the Gaussian BC to receivers in $\mathcal{S}$ (assuming no cache memories).

Proposition 5: For fixed cache memory sizes $\mathsf{M}_1,\ldots,\mathsf{M}_K$,

$$\mathsf{C}(\mathsf{M}_1,\ldots,\mathsf{M}_K) \leq \min_{\mathcal{S} \subseteq \{1,\ldots,K\}} R_{\mathcal{S}}^\star.$$

Proof: Omitted. ∎

The proposition immediately gives an implicit upper bound on the global capacity-memory tradeoff.

Theorem 6:

$$\mathsf{C}^\star(\mathsf{M}) \leq \max_{\substack{\mathsf{M}_1,\ldots,\mathsf{M}_K>0:\\ \sum_{k=1}^K \mathsf{M}_k \leq \mathsf{M}}} \min_{\mathcal{S} \subseteq \{1,\ldots,K\}} R_{\mathcal{S}}^\star. \tag{23}$$

D. Exact Results and Comparisons

1) Exact Results: When M is sufficiently large, bound (21b) coincides with the lower bound in Theorem 3.

Corollary 6.1:

$$\mathsf{C}^\star(\mathsf{M}) = \frac{\sum_{k=1}^K \mathsf{C}_k}{K} + \frac{\mathsf{M}}{KD}, \qquad \frac{\mathsf{M}}{D} \geq (K-1)\sum_{k=1}^K \mathsf{C}_k. \tag{24}$$

Proof: Follows from the upper bound in (21b) and the lower bound in (19) specialized to $t = K-1$. ∎

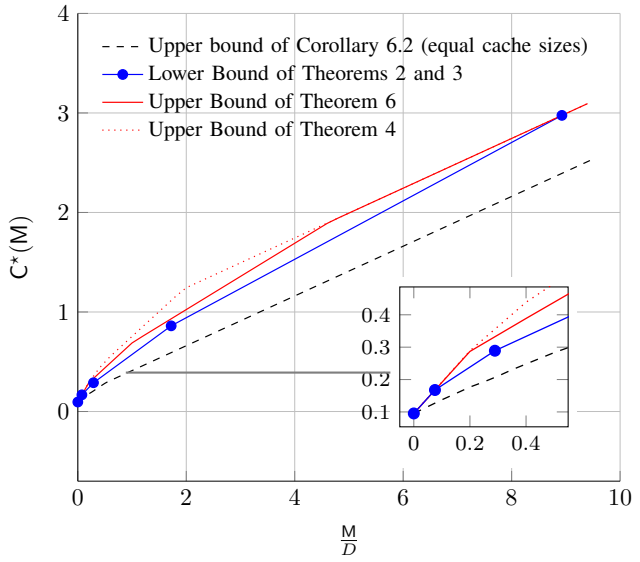

Fig. 1. Bounds on $\mathsf{C}^\star(\mathsf{M})$ for a 4-user Gaussian BC with $P = 1$, $\sigma_1^2 = 4$, $\sigma_2^2 = 2$, $\sigma_3^2 = 0.5$, $\sigma_4^2 = 0.1$.

2) Upper Bounds on the Capacity-Memory Trade-off for Equal Cache Assignment: For comparison, consider a scenario where the total cache memory size M is assigned uniformly among all users, irrespective of their channel statistics:

$$\mathsf{M}_1 = \mathsf{M}_2 = \cdots = \mathsf{M}_K = \frac{\mathsf{M}}{K}. \tag{25}$$

From Proposition 5 we obtain:

Corollary 6.2: Under equal cache assignment, a rate R is achievable only if the tuple

$$\left(R - \frac{\mathsf{M}}{KD}, \ldots, R - \frac{k\mathsf{M}}{K(D-k+1)}, \ldots, R - \frac{\mathsf{M}}{D-K+1}\right)$$

lies in the capacity region without cache memories of the Gaussian BC to receivers $1, \ldots, K$.

3) The Benefits of Cache-Assignment: Figure 1 shows our upper and lower bounds on the global capacity-memory tradeoff $\mathsf{C}^\star(\mathsf{M})$ for a 4-user Gaussian BC. Theorem 6 is only evaluated for the marked points. For comparison we have also plotted an upper bound on the capacity-memory tradeoff $\mathsf{C}(\mathsf{M}/K, \mathsf{M}/K, \ldots, \mathsf{M}/K)$ when all receivers have equal cache memory. We make the following observations:

- Our upper and lower bounds on $\mathsf{C}^\star(\mathsf{M})$ are close and coincide for sufficiently large cache memory M.
- Optimizing the cache assignment significantly increases the capacity-memory tradeoff compared to the upper bound on the capacity memory tradeoff for any scheme with uniform cache assignment.
- The second left-most rate-memory point on the blue line is achieved by assigning all cache memory to the weakest receiver. Given how close our upper and lower bounds lie in the low-cache memory regime, we conclude that in this regime all cache memory should be assigned to the weakest receiver. All other rate-memory points on the blue-line are achieved by assigning positive cache memories to all receivers.

IV. Superposition Piggyback-Coding

We generalize the piggyback coding for erasure BCs in [10, 12] to Gaussian BCs by introducing superposition coding. We piggyback information of multiple stronger receivers on that of a single weak receiver. The scheme is interesting when a receiver is strictly weaker than the others.

Preliminaries: Define

$$\gamma_1 := \frac{\mathsf{C}_{\{2,\ldots,K\}}}{\mathsf{C}_{\{2,\ldots,K\}} + \mathsf{C}_1} \quad \text{and} \quad \gamma_2 := 1 - \gamma_1. \tag{26}$$

The delivery-phase communication takes place in two sub-phases. Sub-phase 1 comprises the first $\lfloor \gamma_1 n \rfloor$ and sub-phase 2 the last $\lfloor \gamma_2 n \rfloor$ channel uses.

Let $\epsilon > 0$ be arbitrarily small, and define the rates

$$\begin{aligned} R^{(\mathrm{A})} &:= \gamma_1 \mathsf{C}_1 = \frac{\mathsf{C}_{\{2,\ldots,K\}} \cdot \mathsf{C}_1}{\mathsf{C}_{\{2,\ldots,K\}} + \mathsf{C}_1} \\ R^{(\mathrm{B})} &:= \gamma_1 \left(\frac{1}{2} \log \left(1 + \frac{(\tilde{\beta}_1 + \tilde{\beta}_2)P}{\sigma_2^2 + \sum_{i=3}^K \tilde{\beta}_i P} \right) - \mathsf{C}_1 \right). \end{aligned} \tag{27}$$

The total rate of the messages is

$$\begin{aligned} R &:= R^{(\mathrm{A})} + R^{(\mathrm{B})} \\ &= \frac{\mathsf{C}_{\{2,\ldots,K\}}}{\mathsf{C}_{\{2,\ldots,K\}} + \mathsf{C}_1} \left(\frac{1}{2} \log \left(1 + \frac{(\tilde{\beta}_1 + \tilde{\beta}_2)P}{\sigma_2^2 + \sum_{i=3}^K \tilde{\beta}_i P} \right) \right). \end{aligned} \tag{28}$$

Allocate cache size

$$\begin{aligned} \mathsf{M}_1 &:= D \cdot R^{(\mathrm{B})} \\ &= \frac{D\mathsf{C}_{\{2,\ldots,K\}}}{\mathsf{C}_{\{2,\ldots,K\}} + \mathsf{C}_1} \left(\frac{1}{2} \log \left(1 + \frac{(\tilde{\beta}_1 + \tilde{\beta}_2)P}{\sigma_2^2 + \sum_{i=3}^K \tilde{\beta}_i P} \right) - \mathsf{C}_1 \right). \end{aligned} \tag{29}$$

at the weakest receiver 1, and no cache size at the other receivers. The total cache size is thus $\mathsf{M} = \mathsf{M}_1$.

Split each message W_d, $d \in \{1, \ldots, D\}$ into two parts

$$W_d = (W_d^{(\mathrm{A})}, W_d^{(\mathrm{B})}), \tag{30}$$

which are of rates $R^{(\mathrm{A})}$ and $R^{(\mathrm{B})}$.

Code constructions: For the communication in the first sub-phase, we use a Gaussian K-level superposition code $\mathcal{C}_1$, where each codebook of a level k is generated with power $\tilde{\beta}_k P$ and contains

$$\begin{cases} \lfloor 2^{nR^{(\mathrm{A})}} \rfloor \text{ codewords}, & \text{if } k = 1, \\ \lfloor 2^{nR^{(\mathrm{B})}} \rfloor \text{ codewords}, & \text{if } k = 2, \ldots, K. \end{cases} \tag{31}$$

Denote by $x^{\gamma_1 n}(w_K | w_1, \ldots, w_{K-1})$ the codeword in the highest level of this superposition code that corresponds to the message tuple $(w_1, \ldots, w_K)$.

For the communication in the second sub-phase, we construct a code $\mathcal{C}_2$ of length $\lfloor \gamma_2 n \rfloor$ and that achieves rate $\mathsf{C}_{\{2,\ldots,K\}}$ to all receivers $2, \ldots, K$.

Caching Phase: Store all messages $W_1^{(\mathrm{B})}, \ldots, W_D^{(\mathrm{B})}$ in the cache memory of receiver 1. This is possible by (29).

Delivery Phase: Transmission takes place in two sub-phases. In sub-phase 1, which is of length $\lfloor \gamma_1 n \rfloor$, the transmitter sends the codeword

$$x^{n\gamma_1}\big(W^{\mathrm{B}}_{d_K}|W^{(\mathrm{A})}_{d_1}, W^{(\mathrm{B})}_{d_2}, \ldots, W^{(\mathrm{B})}_{d_{K-1}}\big).$$

In sub-phase 2, which is of length $\lfloor \gamma_2 n \rfloor$, the transmitter uses codebook $\mathcal{C}_2$ to send messages $W^{(\mathrm{A})}_{d_2}, \ldots, W^{(\mathrm{A})}_{d_{K-1}}, W^{\mathrm{A}}_{d_K}$.

Decoding is as follows. For $k \in \{2, \ldots, K\}$, Receiver k decodes its desired message parts $W^{(\mathrm{A})}_{d_k}$ and $W^{(\mathrm{B})}_{d_k}$ using standard super-position decoding. I.e., in both sub-phases, each of these receivers decodes all levels up to the level containing its desired message part, while treating higher levels as noise.

Receiver 1 only has to decode $W^{(\mathrm{A})}_{d_1}$, because it can retrieve $W^{(\mathrm{B})}_{d_1}$ directly from its cache memory. To decode $W^{(\mathrm{A})}_{d_1}$ it performs the following steps:

1) It retrieves messages $W^{(\mathrm{B})}_{d_2}, \ldots, W^{(\mathrm{B})}_{d_{K-1}}, W^{(\mathrm{B})}_{d_K}$ from its cache memory.
2) It forms the subcodebook $\mathcal{C}_{1,1} \subseteq \mathcal{C}_1$ that contains all highest-level codewords that are "compatible" with the retrieved messages:

$$\mathcal{C}_{1,1} := \Big\{ x^{n\gamma_1}\big(W^{\mathrm{B}}_{d_K}|w, W^{(\mathrm{B})}_{d_2}, \ldots, W^{(\mathrm{B})}_{d_{K-1}}\big) \Big\}_{w=1}^{2^{nR^{(\mathrm{A})}}} \tag{32}$$

3) It decodes its desired message $W^{(\mathrm{A})}_{d_1}$ by restricting attention to subcodebook $\mathcal{C}_{1,1}$.

Analysis: In sub-phase 2 the probability of decoding error tends to 0 as $n \to \infty$ by the way we constructed codebook $\mathcal{C}_2$ and because $nR^{(\mathrm{A})} \leq \lfloor \gamma_2 n \rfloor \mathsf{C}_{\{2,\ldots,K\}}$.

In sub-phase 1, Receivers $k \in \{2, \ldots, K\}$ can reliably decode their message $W^{(B)}_{d_k}$ because (16) and (27) ensure that

$$nR^{(A)} + nR^{(B)} \leq \lfloor n\gamma_1 \rfloor \frac{1}{2} \log\Big(1 + \frac{(\tilde{\beta}_1 + \tilde{\beta}_2)P}{\sigma_2^2 + \sum_{i=3}^{K} \tilde{\beta}_i P}\Big) \tag{33}$$

$$nR^{(B)} \leq \lfloor n\gamma_1 \rfloor \frac{1}{2} \log\Big(1 + \frac{\tilde{\beta}_k P}{\sigma_k^2 + \sum_{i=k+1}^{K} \tilde{\beta}_i P}\Big). \tag{34}$$

Finally, Receiver 1 can decode with arbitrarily small probability of error because subcodebook $\mathcal{C}_{1,1}$ contains $2^{nR^{(\mathrm{A})}}$ Gaussian codewords of average power P and $nR^{(\mathrm{A})} \leq \lfloor n\gamma_1 \rfloor \mathsf{C}_1$.

V. Multi-Piggyback Coding

Next, we generalize the Maddah-Ali & Niesen coded caching scheme of [1] to receivers with unequal cache sizes and we combine it with the piggyback coding idea from [9, 10] to account for different channel conditions at the various receivers.

The scheme is parametrized by an integer number $t \in \{1, \ldots, K\}$, which indicates in how many cache memories each message-part is stored.

Preliminaries: Split each message W_d into $\tau^{(t)}$ independent submessages:

$$W_d = \big\{ W_{d,\mathcal{G}^{(t)}_\ell} : \quad \ell = 1, \ldots, \tau^{(t)} \big\},$$

with each message $W_{d,\mathcal{G}^{(t)}_\ell}$ being of rate[3]

$$R_\ell := \frac{1}{\mu^{(t)}} \prod_{k \in \bar{\mathcal{G}}^{(t)}_\ell} \mathsf{C}_k. \tag{35}$$

The total message rate is

$$R := \sum_{\ell=1}^{\tau^{(t)}} R_\ell = \frac{1}{\mu^{(t)}} \sum_{\ell=1}^{\tau^{(t)}} \prod_{k \in \bar{\mathcal{G}}^{(t)}_\ell} \mathsf{C}_k. \tag{36}$$

Assign to receiver $k \in \{1, \ldots, K\}$ a cache memory of size

$$\mathsf{M}_k = D \sum_{\{\ell:\ k \in \mathcal{G}^{(t)}_\ell\}} R_\ell. \tag{37}$$

Caching Phase: For each $d \in \{1, \ldots, D\}$, store the tuple

$$\Big\{ W_{d,\mathcal{G}^{(t)}_\ell} : \quad k \in \mathcal{G}^{(t)}_\ell,\ \ell = 1, \ldots, \tau^{(t)} \Big\} \tag{38}$$

in the cache memory of Receiver $k \in \{1, \ldots, K\}$. This is possible by the choice of M_k in (37).

Delivery Phase: If $t = K$, there is nothing to send in the delivery phase. We thus assume in the following that $t < K$.

Transmission takes place in $N^{(t)}$ periods, where each Period $j \in \{1, \ldots, N^{(t)}\}$ consists of

$$n_j = \left\lfloor n \cdot \frac{\prod_{k \in \bar{\mathcal{S}}^{(t)}_j} \mathsf{C}_k}{\mu^{(t)}} \right\rfloor \tag{39}$$

consecutive channel uses.

For the transmission in Period j, we construct a power-P Gaussian codebook $\mathcal{C}_j$ of length n_j and rate

$$R_{\mathrm{per}-j} = \left(\max_{k' \in \mathcal{S}^{(t)}_j} \mathsf{C}_{k'} \right) \cdot \frac{1}{\mu^{(t)}} \prod_{k \in \bar{\mathcal{S}}^{(t)}_j} \mathsf{C}_k. \tag{40}$$

The transmitter computes the channel inputs for period $j \in \{1, \ldots, N\}$, as a function of the messages

$$\Big\{ W_{d_k, \mathcal{S}^{(t)}_j \setminus \{k\}} : \quad \forall k \in \mathcal{S}^{(t)}_j \Big\} \tag{41}$$

as follows.

1) It zero-pads the binary representation of each message in (41) to the same length $n_j R_{\mathrm{per}-j}$ bits, and creates the XOR of these zero-padded messages:

$$\bar{W}_{\mathrm{XOR}, \mathcal{S}^{(t)}_j} = \bigoplus_{k \in \mathcal{S}^{(t)}_j} \bar{W}_{d_k, \mathcal{S}^{(t)}_j \setminus \{k\}}, \tag{42}$$

where $\bar{W}_{d_k, \mathcal{S}^{(t)}_j \setminus \{k\}}$ denotes the zero-padded version of message $W_{d_k, \mathcal{S}^{(t)}_j \setminus \{k\}}$.

2) It encodes the XOR message $\bar{W}_{\mathrm{XOR}, \mathcal{S}^{(t)}_j}$ using codebook $\mathcal{C}_j$, and sends the resulting codeword over the channel during period j.

[3]To be precise, to ensure that the probability of error of our scheme tends to 0 as $n \to \infty$ the rate should be slightly smaller than what is indicated in (35). We ignore this technicality for ease of exposition.

Each Receiver $k \in \{1,\dots,K\}$ can retrieve submessages

$$\Big\{W_{d_k,\mathcal{G}_\ell^{(t)}}\colon \quad k\in\mathcal{G}_\ell^{(t)},\ \ell=1,\dots,\tau^{(t)}\Big\} \tag{43}$$

directly from its cache, see (38). It thus only has to decode the remaining submessages of W_{d_k}.

Receiver k observes the channel outputs $Y_k^n = (Y_{k,1},\dots,Y_{k,n})$, which it decomposes into $N^{(t)}$ subsequences of outputs observed at each period $j \in \{1,\dots,N^{(t)}\}$:

$$Y_{k,\text{per}-j}^{n_j} := \big(Y_{k,\sum_{j'=1}^{j-1} n_{j'}+1},\dots,Y_{k,\sum_{j'=1}^{j} n_{j'}}\big).$$

For every $j \in \{1,\dots,N^{(t)}\}$ so that $k \in \mathcal{S}_j^{(t)}$, Receiver k performs the following steps to decode message $W_{d_k,\mathcal{S}_j^{(t)}\setminus\{k\}}$:

1) It retrieves messages

$$\Big\{W_{d_{k'},\mathcal{S}_j\setminus\{k'\}}\colon \quad \forall k'\in\mathcal{S}_j^{(t)}\setminus\{k\}\Big\} \tag{44}$$

from its cache memory, and pads their binary representations to the same maximum length $n_j R_{\text{per}-j}$ bits. So, it computes the tuple

$$\Big\{\bar{W}_{d_{k'},\mathcal{S}_j^{(t)}\setminus\{k'\}}\colon \quad \forall k'\in\mathcal{S}_j^{(t)}\setminus\{k\}\Big\}. \tag{45}$$

2) It extracts a subcodebook $\mathcal{C}_{j,k}$ from $\mathcal{C}_j$ by restricting to codewords that are "compatible" with the zero-padded submessages in (45). Let $R_{j,k}$ be the rate of the desired submessage $W_{d_k,\mathcal{S}_j^{(t)}\setminus\{k\}}$, and let $\bar{\mathcal{W}}_{j,k}$ denote the set of the binary representations of $\{1,\dots,2^{n_j R_{j,k}}\}$ zero-padded to length $n_j R_{\text{per}-j}$. Then,

$$\mathcal{C}_{j,k} := \Big\{x_j^{n_j}\Big(\bar{w}\bigoplus_{k'\in\mathcal{S}_j^{(t)}\setminus\{k\}}\bar{W}_{d_{k'},\mathcal{S}_j^{(t)}\setminus\{k'\}}\Big)\Big\}_{\bar{w}\in\bar{\mathcal{W}}_{j,k}}.$$

3) It decodes the XOR message $\bar{W}_{\text{XOR},\mathcal{S}_j}$ using the restricted codebook $\mathcal{C}_{j,k}$.
4) From $\hat{\bar{w}}_{\text{XOR},\mathcal{S}_j}$ and the messages in (45), it forms

$$\hat{\bar{w}}_{d_k,\mathcal{S}_j^{(t)}\setminus\{k\}} = \hat{\bar{w}}_{\text{XOR},\mathcal{S}_j}\bigoplus_{k'\in\mathcal{S}_j^{(t)}\setminus\{k\}}\bar{W}_{d_{k'},\mathcal{S}_j^{(t)}\setminus\{k'\}}, \tag{46}$$

and retrieves the zero-padding from $\hat{\bar{w}}_{d_k,\mathcal{S}_j^{(t)}\setminus\{k\}}$ to obtain $\bar{w}_{d_k,\mathcal{S}_j^{(t)}\setminus\{k\}}$.

After the last period $N^{(t)}$, each Receiver k produces the estimate $\hat{W}_{d_k}$ that corresponds to the retrieved tuple in (43) and to the decoded tuple

$$\Big\{\hat{w}_{d_k,\mathcal{G}_\ell^{(t)}}\colon\ k\notin\mathcal{G}_\ell^{(t)},\ \ell=1,\dots,\tau^{(t)}\Big\}, \tag{47}$$

and declares this as its estimate of message W_{d_k}.

Analysis: When every padded XOR-message $\bar{W}_{\text{XOR},\mathcal{S}_j}$ is decoded correctly by all its intended receivers, then all receivers $1,\dots,K$ produce the correct estimate of their desired messages $W_{d_1},\dots,W_{d_K}$.

Fix $j \in \{1,\dots,N^{(t)}\}$ and consider the probability of decoding error of $\bar{W}_{\text{XOR},\mathcal{S}_j}$ at a specific receiver $k \in \mathcal{S}_j$. Let ℓ be such that

$$\mathcal{G}_\ell^{(t)} = \mathcal{S}_j^{(t)}\setminus\{k\}.$$

The probability of the considered decoding error tends to 0 as n (and thus n_j) tends to ∞ because the subcodebook $\mathcal{C}_{j,k}$ contains 2^{nR_ℓ} Gaussian codewords and because $\frac{nR_\ell}{n_j} = \mathsf{C}_k$, see (35) and (39).

VI. Conclusion

We studied K-user Gaussian broadcast channels with receiver cache assignment and established lower and upper bounds on their global capacity-memory tradeoffs. We proposed a joint cache-channel coding scheme that can benefit from unequal cache sizes at the receivers and account for different channel qualities. By carefully assigning larger cache memories to weaker receivers we can achieve rates that are impossible to achieve with equal cache assignment.

References

[1] M. A. Maddah-Ali, U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May, 2014.

[2] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *IET Commun.*, 2016, Vol. 10, Iss. 17, pp. 2315–2318.

[3] M. M. Amiri and D. Gündüz, "Fundamental limits of caching: improved delivery rate-cache capacity trade-off," *IEEE Trans. on Comm.*, vol. 65, no. 2, pp. 806–815, Feb. 2017.

[4] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," in in Proc. of *IEEE ISIT*, Barcelona, July, 2016.

[5] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," in Proc. of *IEEE ISIT*, Hong Kong, June, 2015.

[6] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in Proc. of *IEEE ISIT*, Hong Kong, June, 2015.

[7] C.-Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," *arXiv*, 1601.05690.

[8] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv*, 1502.03124, Feb., 2015.

[9] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in Proc. of *IEEE ISWCS*, Bruxelles, Aug., 2015.

[10] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy Broadcast Networks with Receiver Caching," submitted to *IEEE Trans. Inf. Theory*, *arXiv*, 1605.02317.

[11] A. S. Cacciapuoti, M. Caleffi, M. Ji, J. Llorca, A. M. Tulino,"Speeding up future video distribution via channel-aware caching-aided coded multicast," *IEEE JSAC in Comm.*, vol. 34, no. 8, Aug., 2016.

[12] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "An Upper Bound on the Capacity-Memory Tradeoff of Degraded Broadcast Channels," in Proc. of *IEEE ISTC*, Brest, Sep., 2016.

[13] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv*, 1504.01452.

[14] P. Hassanzadeh, E. Erkip, J. Llorca, and A. Tulino, "Distortion-memory tradeoffs in Cache-aided wireless video delivery," in *Allerton Conf. on Comm., Control and Comp.*, Monticello (IL), Oct., 2015.

[15] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," in *Allerton Conf. on Comm., Control and Comp.*, Monticello (IL), Oct., 2015.

[16] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," *arXiv*, 1511.03961.

[17] M. M. Amiri, Q. Yang, D. Gndz, "Decentralized Coded Caching with Distinct Cache Capacities," *arXiv*, 1610.03792, Oct., 2016.

[18] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *arXiv*, 1504.01123v1.

[19] M. Wigger, R. Timo, and S. Shamai, "Complete Interference Mitigation through Receiver-Caching in Wyner's Networks," in *Inf. Theory Workshop*, Cambridge, Sep., 2016.

[20] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1469 –1482, April, 2006.