Benefits of Cache Assignment on Degraded Broadcast Channels

[†]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*—Degraded K-receiver broadcast channels (BC) are studied when receivers have cache memories. Lower and upper bounds are derived on the *capacity-memory tradeoff*, i.e., on the largest rate that can be achieved over the BC as a function of the receivers' cache sizes. The lower bounds are achieved by two new coding schemes that benefit from non-uniform cache assignment. In some special cases, the lower and upper bounds coincide. The paper also provides lower and upper bounds on the *global capacity-memory tradeoff* of degraded BCs, i.e., on the largest capacity-memory tradeoff that can be attained by optimizing the receivers cache-assignment subject to a *total* cache memory budget. The bounds coincide when the total cache memory budget is sufficiently small or sufficiently large, with thresholds depending on the BC statistics. For a small total cache budget M, it is optimal to assign all the cache memory to the weakest receiver. In this regime, the global capacity-memory tradeoff grows as $\frac{\mathsf{M}}{D}$, where D denotes the total number of files in the system. For a large total cache budget, it is optimal to assign a positive cache memory to every receiver, where weaker receivers are assigned larger cache memories than stronger receivers. When the total cache budget M exceeds a threshold, then the global capacity-memory tradeoff grows as $\frac{1}{K} \cdot \frac{\mathsf{M}}{D}$. A uniform cache-assignment policy is suboptimal.

I. Introduction

This paper studies the degraded broadcast channel (BC) when the receivers are equipped with cache memories and can store (i.e., cache) contents before the actual communication phase (the so called delivery phase). Our previous work in [2]–[4] on asymmetric erasure BCs with caching receivers has shown that assigning larger cache memories to the weaker receivers significantly improves the performance compared to the traditional uniform cache assignment that is studied in [1], [6]–[13]. In addition to mitigating the rate-bottleneck of the network, a non-uniform cache assignment creates new coding opportunies by *joint cache-channel coding* [2]–[5]. We continue the sprit embodied in the previous work, i.e., the benefit brought upon by non-uniform cache sizes, and generalize it in the channel model, coding techniques, and bounds.

The quantity of interest in this paper is the *capacity-memory tradeoff*, i.e., the largest rate R permitting reliable communication as a function of the cache sizes. We provide upper and lower bounds on the capacity-memory tradeoff for degraded BCs. The lower bounds are obtained by two new coding schemes that exploit unequal cache assignments and asymmetric channel conditions. More specifically, the first scheme combines the piggyback-coding idea in [2] with superposition coding, and the second combines it with Maddah-Ali and Niesen's coded caching [1].

The new upper bound improves on the existing ones in [3, 14]. It coincides with the superposition piggyback-coding lower bound when only the weakest receiver is provided with a cache memory and the size is below a certain threshold (which depends on the BC statistics). It coincides with the generalized coded-caching lower bound for a particular cache assignment that assigns larger cache memories to the weaker receivers.

The proposed upper and lower bounds suggest that in a good cache assignment, weaker receivers are provided with larger cache memories compared to the stronger receivers. To make this statement more precise, we derive upper and lower bounds on the *global capacity-memory tradeoff*, where the cache assignment is optimized subject to a *total cache constraint*. The bounds coincide for small and large total cache memories. The lower bound is achieved by assigning all the cache memory to the weakest receiver when the total cache memory is small, and by distributing it among all the receivers when its size exceeds some threshold.

Numerical evaluations of the new upper bound confirm that uniform cache assignment is suboptimal in all regimes.

II. Problem Definition

Consider a transmitter and receivers $1, \ldots, K$. The transmitter has access to a library with D independent messages, $W_1, \ldots, W_D$, each distributed uniformly over the set $\{1, \ldots, \lfloor 2^{nR} \rfloor\}$. I.e. $R \geq 0$ denotes the rate of transmission and n is the transmission blocklength. In this work, we assume that there are more messages than receivers, $D \geq K$.

Each receiver $k \in \mathcal{K} := \{1, \ldots, K\}$ is equipped with a cache of size $\mathsf{M}_k \geq 0$. Communication takes place in two phases. For the first so called placement phase, the transmitter chooses caching functions $\{g_k : \{1, \ldots, \lfloor 2^{nR} \rfloor\}^D \to \{1, \ldots, \lfloor 2^{n\mathsf{M}_k} \rfloor\}\}_{k=1}^K$ and places

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

in receiver k's cache.

The subsequent delivery phase takes place over a *degraded* BC [21] with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}_1, \ldots, \mathcal{Y}_K$, and the channel transition law $\Gamma(y_1, \ldots, y_K|x)$ that decomposes as follows for all $x \in \mathcal{X}$, $y_1 \in \mathcal{Y}_1, \ldots, y_K \in \mathcal{Y}_K$:

$$\Gamma(y_1, \ldots, y_K|x) = \Gamma_K(y_K|x) \cdot \Gamma_{K-1}(y_{K_1}|y_K) \cdots \Gamma_1(y_1|y_2).$$

Without loss of generality, we order the receivers $1, \ldots, K$ from the weakest to the strongest.

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

$$X^n = f_{\mathbf{d}}(W_1, \ldots, W_D), \tag{2}$$

using the encoding function $f_{\mathbf{d}} : \{1, \ldots, \lfloor 2^{nR} \rfloor\}^D \to \mathcal{X}^n$.

Receiver $k \in \mathcal{K}$ observes the channel output sequence $Y_k^n = (Y_{k,1}, \ldots, Y_{k,n})$. Given the demand vector $\mathbf{d}$, the cache content $\mathbb{V}_k$, and the channel output Y_k^n, it produces its estimate of the desired message W_{d_k} as

$$\hat{W}_k := \varphi_{k,\mathbf{d}}(Y_k^n, \mathbb{V}_k), \tag{3}$$

where $\varphi_{k,\mathbf{d}} : \mathcal{Y}_k^n \times \{1, \ldots, \lfloor 2^{n\mathsf{M}_k} \rfloor\} \to \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ is the decoding function.

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

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

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

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

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

Our main goal in this paper 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

$$\sum_{k=1}^{K} \mathsf{M}_k \leq \mathsf{M}. \tag{4}$$

Definition 2: The *global capacity-memory tradeoff* is

$$\mathsf{C}^\star(\mathsf{M}) := \max_{\substack{\mathsf{M}_1, \ldots, \mathsf{M}_K > 0: \\ \sum_{k=1}^{K} \mathsf{M}_k \leq \mathsf{M}}} \mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K). \tag{5}$$

Remark 1: The global capacity memory tradeoff depends on the BC law $\Gamma(y_1, \ldots, y_K|x)$ only through its marginal conditional laws. All our results thus also apply to *stochastically degraded BCs* [21].

Without cache memories, i.e., $\mathsf{M}_1 = \ldots = \mathsf{M}_2 = 0$, we have

$$\mathsf{C}(\mathsf{M}_1 = 0, \ldots, \mathsf{M}_K = 0) = \mathsf{C}_{\mathcal{K}} \tag{6}$$

where $\mathsf{C}_{\mathcal{K}}$ is (see [21]) :

$$\mathsf{C}_{\mathcal{K}} := \max \min \big\{ I(U_1; Y_1),\ I(U_2; Y_2|U_1), \\ \ldots, I(U_{K-1}; Y_{K-1}|U_{K-2}), I(X; Y_K|U_{K-1}) \big\}. \tag{7}$$

The maximization in (7) is over all auxiliary random variables $U_1, \ldots, U_{K-1}, X, Y_1, \ldots, Y_K$ that satisfy

$$P_{Y_1 \cdots Y_K|X}(y_1, \ldots, y_K|x) = \Gamma(_1, \ldots, y_k|x) \tag{8}$$

and form the Markov chain $U_1 - \cdots - U_{K-1} - X - (Y_1, \ldots, Y_K)$.

III. Results on the Capacity-Memory Tradeoff

A. Upper Bound on Capacity-Memory Tradeoff

The upper bound is formulated in terms of the following parameters. For each receiver set

$$\mathcal{S} = \{j_1, \ldots, j_{|\mathcal{S}|}\} \subseteq \mathcal{K}, \qquad j_1 < \cdots < j_{|\mathcal{S}|}, \tag{9}$$

define

$$\alpha^\star_{\mathcal{S},1} := \frac{\mathsf{M}_{j_1}}{D} \tag{10a}$$

and for $k \in \{2, \ldots, |\mathcal{S}|\}$:

$$\alpha^\star_{\mathcal{S},k} := \min \Bigg\{ \frac{\sum_{i=1}^{k} \mathsf{M}_{j_i}}{D - k + 1}, \\ \frac{1}{|\mathcal{S}| - k + 1} \Bigg(\frac{|\mathcal{S}|}{D} \sum_{i=1}^{|\mathcal{S}|} \mathsf{M}_{j_i} - \sum_{i=1}^{k-1} \alpha^\star_{\mathcal{S},i} \Bigg) \Bigg\}. \tag{10b}$$

Theorem 1: There exist random variables $X, Y_1, \ldots, Y_K$ and for *every* receiver set $\mathcal{S}$ as in (9) random variables $\{U_{\mathcal{S},1}, \ldots, U_{\mathcal{S},|\mathcal{S}|-1}\}$, such that (8) and the Markov chain

$$U_{\mathcal{S},1} - U_{\mathcal{S},2} - U_{\mathcal{S},|\mathcal{S}|} - \cdots - U_{\mathcal{S},|\mathcal{S}|-1} - X - \big(Y_{j_1}, \ldots, Y_{j_{|\mathcal{S}|}}\big) \tag{11}$$

are satisfied and the following inequalities hold for all $\mathcal{S}$:

$$\mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K) \leq I\big(U_{\mathcal{S},k}; Y_{j_k} | U_{\mathcal{S},k-1}\big) + \alpha^\star_{\mathcal{S},k}, \\ \forall k \in \{1, \ldots, |\mathcal{S}| - 1\}, \tag{12a}$$

$$\mathsf{C}(\mathsf{M}_1, \ldots, \mathsf{M}_K) \leq I\big(X; Y_{j_{|\mathcal{S}|}} | U_{\mathcal{S},|\mathcal{S}|-1}\big) + \alpha^\star_{\mathcal{S},|\mathcal{S}|}. \tag{12b}$$

Proof: Omitted. See [23]. ∎

The converse in Theorem 1 is weakened if the constraints in (12) are relaxed for certain receiver sets $\mathcal{S}$, or if in these constraints the input/output random variables $X, Y_{j_1}, \ldots, Y_{j_{|\mathcal{S}|}}$ are allowed to depend on the receiver set $\mathcal{S}$. For this latter relaxation, Theorem 1 results in the following corollary.

Corollary 2: The rate-memory tuple $(R, \mathsf{M}_1, \ldots, \mathsf{M}_K)$ is achievable only if for every receiver set $\mathcal{S} \subseteq \mathcal{K}$:

$$\Big(R - \alpha^\star_{\mathcal{S},1}, R - \alpha^\star_{\mathcal{S},2}, \ldots, R - \alpha^\star_{\mathcal{S},|\mathcal{S}|} \Big) \in \mathcal{C}_{\mathcal{S}}, \tag{13}$$

where $\mathcal{C}_{\mathcal{S}}$ denotes the capacity region of the degraded BC with the receivers in $\mathcal{S}$ (disregarding the receivers in $\mathcal{K} \backslash \mathcal{S}$) when there are no cache memories [21].

Remark 2: The upper bounds in Theorem 1 and Corollary 2 can be made looser by replacing the parameters $\{\alpha^\star_{\mathcal{S},k}\}$ by

$$\tilde{\alpha}_{\mathcal{S},k} = \frac{\sum_{i=1}^{k} \mathsf{M}_{j_i}}{D - k + 1}. \tag{14}$$

The same is true also if they are replaced by

$$\alpha'_{\mathcal{S},k} = \frac{1}{D}\sum_{i=1}^{|\mathcal{S}|} \mathsf{M}_{j_i}. \tag{15}$$

Replacing the parameters $\{\alpha^\star_{\mathcal{S},k}\}$ by the parameters $\{\alpha'_{\mathcal{S},k}\}$ in Corollary 2, we recover the previous upper bounds in [3, Theorem 9] and [14, Theorem 1].

B. Lower Bounds on the Capacity-Memory Tradeoff

We start with a general lower bound that simply exploits the local caching gain. Similar to [20, Proposition 1], we have:

Proposition 3 (Local caching gain): For all $\Delta > 0$:

$$\mathsf{C}(\mathsf{M}_1+\Delta,\ldots,\mathsf{M}_K+\Delta) \geq \mathsf{C}(\mathsf{M}_1,\ldots,\mathsf{M}_K) + \frac{\Delta}{D}. \tag{16}$$

Proof: The lower bound is achieved by storing a rate-$\frac{\Delta}{D}$ submessage of every message of the library in the cache memory of every receiver. These submessages can be retrieved locally and thus not be transmitted in the delivery phase. ■

We next present two lower bounds on the capacity-memory tradeoff based on the coding schemes sketched in Sections V and VI. The first assigns a cache memory only to the weakest receiver and the second assigns a cache memory to every receivers, but such that weaker receivers are assigned larger cache memories compared to the stronger receivers.

Let $(U_1^\star,\ldots,U_{K-1}^\star,X^\star)$ be a K-tuple of random variables that achieves the symmetric-capacity C_0; i.e., it is a solution to the optimization problem in (7). Define

$$\mathsf{M}_1^{\mathsf{s}} := \frac{D}{K-1}\big(I(U_1^\star;Y_2) - I(U_1^\star;Y_1)\big). \tag{17}$$

Theorem 4 (Superposition Piggyback Coding): When $\mathsf{M}_1 \leq \mathsf{M}_1^{\mathsf{s}}$, then irrespective of cache sizes $\mathsf{M}_2,\ldots,\mathsf{M}_K$ we have

$$\mathsf{C}(\mathsf{M}_1,\ldots,\mathsf{M}_K) \geq \mathsf{C}_{\mathcal{K}} + \frac{\mathsf{M}_1}{D}, \qquad \mathsf{M}_1 \leq \mathsf{M}_1^{\mathsf{s}}. \tag{18}$$

Proof: Achieved by the scheme in Section V, see also [23]. ■

For each $t \in \mathcal{K}$, let

$$\mathcal{G}_1^{(t)},\ldots,\mathcal{G}_{\binom{K}{t}}^{(t)} \tag{19}$$

denote all the unordered subsets of $\{1,\ldots,K\}$ that are of size-t. Also, let $\mathcal{G}_\ell^{(t),c} := \{1,\ldots,K\}\backslash\mathcal{G}_\ell^{(t)}$ for all $\ell \in \{1,\ldots,\binom{K}{t}\}$. For a given input distribution P_X, define for each $t \in \{1,\ldots,K-1\}$ the memory sizes

$$\mathsf{M}_k^{(t)} := D\cdot\frac{\sum_{\left\{\ell:\ k\in\mathcal{G}_\ell^{(t)}\right\}}\prod_{k'\in\mathcal{G}_\ell^{(t),c}} I(X;Y_{k'})}{\sum_{j=1}^{\binom{K}{t+1}}\prod_{k'\in\mathcal{G}_j^{(t+1),c}} I(X;Y_{k'})}, \quad k\in\mathcal{K}, \tag{20a}$$

where the denominator is defined to be 1 when $t = K-1$. Also, for each $t \in \{1,\ldots,K-1\}$, define the transmission rate

$$R^{(t)} := \frac{\sum_{\ell=1}^{\binom{K}{t}}\prod_{k'\in\mathcal{G}_\ell^{(t),c}} I(X;Y_{k'})}{\sum_{j=1}^{\binom{K}{t+1}}\prod_{k'\in\mathcal{G}_j^{(t+1),c}} I(X;Y_{k'})}, \tag{20b}$$

where the denominator is again 1 for $t = K-1$.

Theorem 5 (Generalized Coded Caching): Fix an input distribution P_X. For each $t \in \{1,\ldots,K-1\}$:

$$\mathsf{C}\Big(\mathsf{M}_1^{(t)},\ldots,\mathsf{M}_K^{(t)}\Big) \geq R^{(t)}, \tag{21}$$

where $\mathsf{M}_1^{(t)},\ldots,\mathsf{M}_K^{(t)}$ and $R^{(t)}$ are defined by (20) and the chosen input distribution P_X.

Proof: The lower bound is achieved by the scheme outlined in Section VI, see [23]. ■

C. Exact Results

The upper and lower bounds match in two special cases.

Proposition 6 (Small Cache Memory at the Weakest Receiver): Suppose $M_1 > 0$ and $\mathsf{M}_2 = \ldots = \mathsf{M}_K = 0$. Then,

$$\mathsf{C}(\mathsf{M}_1,0,\ldots,0) = \mathsf{C}_{\mathcal{K}} + \frac{\mathsf{M}_1}{D}, \qquad \mathsf{M}_1 \leq \mathsf{M}_1^{\mathsf{s}}, \tag{22}$$

where $\mathsf{M}_1^{\mathsf{s}}$ is defined in (17).

Proof: The achievability is by superposition piggyback coding, see Theorem 4, and the converse is by Corollary 2, where one only considers $\mathcal{S} = \mathcal{K}$. ■

The performance in (22) corresponds to a *perfect* global caching gain where each and every receiver can benefit from receiver 1's *cache content* as if it was locally present.

Proposition 7 (Large Cache Memories): For each $k \in \mathcal{K}$, let $\mathsf{M}_k^{\star(K-1)}$ be given by (20a) when P_X is a maximizer of

$$\mathsf{C}_{\text{Avg}} := \frac{1}{K}\cdot\max_{P_X}\sum_{k=1}^{K} I(X;Y_k). \tag{23}$$

For any $\Delta \geq 0$:

$$\begin{aligned}\mathsf{C}\Big(&\mathsf{M}_1^{\star(K-1)}+\Delta,\ \ldots,\ \mathsf{M}_K^{\star(K-1)}+\Delta\Big)\\ &= \mathsf{C}_{\text{Avg}} + \frac{\sum_{k=1}^{K}\mathsf{M}_k^{\star(K-1)}}{K\cdot D} + \frac{\Delta}{D}.\end{aligned} \tag{24}$$

Proof: For $\Delta = 0$, the achievability is by the generalized coded caching with $t = K-1$, see Theorem 5. For $\Delta > 0$, it follows from the case $\Delta = 0$ and from Proposition 3. The converse follows from Theorem 1 for $\mathcal{S} = \{k\}$, $k \in \mathcal{K}$, by relaxing the parameters $\alpha^\star_{\mathcal{S},k}$ to $\alpha'_{\mathcal{S},k}$. ■

IV. Results on the Global Capacity-Memory Tradeoff

A. Upper Bounds

Theorem 1 directly yields the following result.

Theorem 8 (Upper Bound): There exist random variables $X, Y_1,\ldots,Y_K$ and for *every* receiver set $\mathcal{S}$ as in (9) random variables $\{U_{\mathcal{S},1},\ldots,U_{\mathcal{S},|\mathcal{S}|-1}\}$, so that (8) and (11) hold, and so that *for some* $\mathsf{M}_1,\ldots,\mathsf{M}_K \geq 0$ summing to M and all $\mathcal{S}$:

$$\begin{aligned}&\mathsf{C}^\star(\mathsf{M}) \leq I\big(U_{\mathcal{S},k};Y_{j_k}|U_{\mathcal{S},k-1}\big) + \alpha^\star_{\mathcal{S},k},\quad k \leq |\mathcal{S}|-1\\ &\mathsf{C}^\star(\mathsf{M}) \leq I\big(X;Y_{j_{|\mathcal{S}|}}|U_{\mathcal{S},|\mathcal{S}|-1}\big) + \alpha^\star_{\mathcal{S},|\mathcal{S}|}\end{aligned} \tag{25}$$

where $\{\alpha^\star_{\mathcal{S},k}\}$ are defined in (10).

Solving this optimization problem numerically is computationally complex. Simpler, albeit looser, upper bounds can be obtained by either ignoring some of the constraints in (25); replacing the parameters $\alpha^{\star}_{\mathcal{S},k}$ by $\tilde{\alpha}_{\mathcal{S},k}$ or $\alpha'_{\mathcal{S},k}$ (see (14),(15)); or allowing $X, Y_{j_1}, \ldots, Y_{j_{\mathcal{S}}}$ in (25) to depend on the set $\mathcal{S}$. We present three simpler bounds that will be used in the sequel.

Corollary 9:

$$\mathsf{C}^{\star}(\mathsf{M}) \leq \mathsf{C}_{\mathcal{K}} + \frac{\mathsf{M}}{D} \tag{26}$$

and

$$\mathsf{C}^{\star}(\mathsf{M}) \leq \mathsf{C}_{\text{avg}} + \frac{1}{K} \cdot \frac{\mathsf{M}}{D}. \tag{27}$$

For each set $\mathcal{G}_{\ell}^{(t)}$ (see (19)), let $\mathsf{C}_{\mathcal{G}_{\ell}^{(t)}}$ denote the largest symmetric rate that can be achieved at the receivers in $\mathcal{G}_{\ell}^{(t)}$ (ignoring the receivers in $\mathcal{K}\backslash\mathcal{G}_{\ell}^{(t)}$) when there are no cache memories.

Corollary 10: For each $t \in \mathcal{K}$:

$$\mathsf{C}^{\star}(\mathsf{M}) \leq \frac{1}{\binom{K}{t}} \sum_{\ell=1}^{\binom{K}{t}} \mathsf{C}_{\mathcal{G}_{\ell}^{(t)}} + \frac{t}{K} \cdot \frac{\mathsf{M}}{D}. \tag{28}$$

B. Lower Bound

Let $\mathsf{M}^{(0)} := 0$ and $\mathsf{M}^{\mathsf{s}} := \mathsf{M}_1^{\mathsf{s}}$ (see (17)), and let $R^{(0)} := \mathsf{C}_{\mathcal{K}}$ and $R^{\mathsf{s}} := \mathsf{C}_{\mathcal{K}} + \frac{\mathsf{M}^{\mathsf{s}}}{D}$. Also, recall $\mathsf{M}_k^{(t)}$ and $R^{(t)}$ from (20) and define $\mathsf{M}^{(t)} := \sum_{k=1}^{K} \mathsf{M}_k^{(t)}$ for $t \in \{1, \ldots, K-1\}$.

Theorem 11 (Lower Bound):

$$\mathsf{C}^{\star}(\mathsf{M}) \geq \text{upp hull}\big\{(R^{(0)}, \mathsf{M}^{(0)}),\ (R^{\mathsf{s}}, \mathsf{M}^{\mathsf{s}}), (R^{(1)}, \mathsf{M}^{(1)}), \ldots,\ (R^{(K-1)}, \mathsf{M}^{(K-1)})\big\}. \tag{29}$$

Proof: The proof follows by Theorems 4 and 5, and uses standard memory-sharing arguments [1, 3]. ■

C. Exact Results

Proposition 12 (Small M): For small total cache size M:

$$\mathsf{C}^{\star}(\mathsf{M}) = \mathsf{C}_{\mathcal{K}} + \frac{\mathsf{M}}{D}, \qquad \mathsf{M} \leq \mathsf{M}^{\mathsf{s}}. \tag{30}$$

Proof: The achievability follows from Theorem 11 and the converse is by inequality (26). ■

Proposition 13 (Large M): For large total cache size

$$\mathsf{M} \geq D \cdot (K-1) \cdot K \cdot \mathsf{C}_{\text{avg}}, \tag{31}$$

the global capacity-memory tradeoff $\mathsf{C}^{\star}(\mathsf{M})$ is

$$\mathsf{C}^{\star}(\mathsf{M}) = \mathsf{C}_{\text{avg}} + \frac{1}{K} \cdot \frac{\mathsf{M}}{D}, \tag{32}$$

where C_{avg} is defined in (23).

Proof: The achievability is by Propositions 3 and 7, and the converse is by inequality (27). See [23]. ■

For small total cache sizes, $\mathsf{C}^{\star}(\mathsf{M})$ grows as $\frac{\mathsf{M}}{D}$. This corresponds to a perfect global caching gain as if each receiver could access all cache contents in the network locally. For large total cache sizes, $\mathsf{C}^{\star}(\mathsf{M})$ grows as $\frac{1}{K} \cdot \frac{\mathsf{M}}{D}$. This corresponds to only a local caching gain under a uniform cache assignment. This is, in this regime, each receiver profits only from its local cache contents.

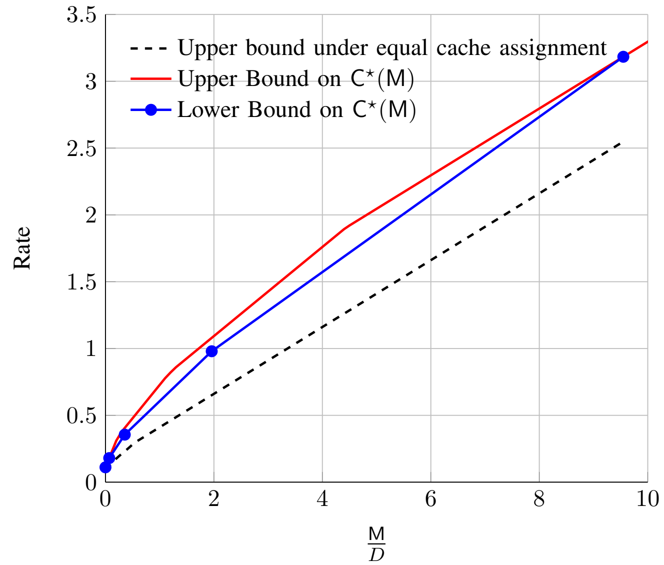

Fig. 1. Bounds on $\mathsf{C}^{\star}(\mathsf{M})$ for a 4-user Gaussian BC with $P=1$, $\sigma_1^2 = 4$, $\sigma_2^2 = 1$, $\sigma_3^2 = 0.5$, $\sigma_4^2 = 0.1$. We have $D = 10$.

Example 1 (4-User Gaussian BC): Consider a $K = 4$ receiver Gaussian BC. At time t, receiver k observes

$$Y_{k,t} = X_t + Z_{k,t}, \tag{33}$$

where X_t is the transmitter's channel input and $\{Z_{k,t}\}$ a sequence of independent and identically distributed centered Gaussian noise with variance $\sigma_k^2 > 0$. We impose an average block-power constraint $P > 0$ in the channel input and order the receivers in increasing strength:

$$\sigma_1^2 = 4\ \geq\ \sigma_2^2 = 1\ \geq\ \sigma_3^2 = 0.5\ \geq\ \sigma_4^2 = 0.1\ > 0.$$

For Gaussian BCs, a zero-mean variance-P Gaussian input distribution P_X maximizes $I(X;Y_k)$ simultaneously for all $k \in \mathcal{K}$ under an input power constraint P. Therefore,

$$\mathsf{C}_{\text{avg}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{2} \log\left(1 + \frac{P}{\sigma_k^2}\right) = \frac{1}{K} \sum_{k=1}^{K} \mathsf{C}_k. \tag{34}$$

Figure 1 shows the upper and lower bounds on $\mathsf{C}^{\star}(\mathsf{M})$ in Theorems 8 and 11. The five blue points indicate the rate-memory points $(R^{(0)}, \mathsf{M}^{(0)})$, $(R^{\mathsf{s}}, M^{\mathsf{s}})$, $(R^{(1)}, \mathsf{M}^{(1)})$, $(R^{(2)}, \mathsf{M}^{(2)})$, and $(R^{(3)}, \mathsf{M}^{(3)})$. For comparison, the figure also shows the upper bound in Theorem 1 for a setup with uniform cache assignment $\frac{\mathsf{M}}{K}$ across all receivers. We observe that a smart cache assignment provides substantial gains in global capacity-memory tradeoff.

In a recent work in [22], we considered Gaussian BCs and derived related, but looser, bounds on $\mathsf{C}^{\star}(\mathsf{M})$. In particular, the bounds in [22] do not coincide for small M.

V. Superposition Piggyback-Coding

We outline the scheme and refer to [23] for details.

Fix a tuple $(U_1^{\star}, \ldots, U_K^{\star}, X^{\star})$ that maximizes (7), and construct a K-level superposition code according to the joint distribution of this tuple. Split $W_d = \left(W_d^{(\mathrm{A})}, W_d^{(\mathrm{B})}\right)$, and cache messages $W_1^{(\mathrm{B})}, \ldots, W_D^{(\mathrm{B})}$ at receiver 1.

The transmitter uses the superposition code to send:

$$\begin{aligned} & W_{d_1}^{(\mathrm{A})}, W_{d_2}^{(\mathrm{B})}, \ldots, W_{d_{K-1}}^{(\mathrm{B})}, W_{d_K}^{(\mathrm{B})} && \text{in level 1} \\ & W_{d_k}^{(\mathrm{A})} && \text{in level } k \in \{2, \ldots, K\}. \end{aligned}$$

Each receiver $k \in \{2, \ldots, K\}$ decodes levels $1, 2, \ldots, k$ using a standard decoder. Receiver 1 only decodes level 1 to which it applies a joint cache-channel decoding rule. It first retrieves messages $W_{d_2}^{(\mathrm{B})}, \ldots, W_{d_K}^{(\mathrm{B})}$ from its cache memory, and extracts a subcodebook $\mathcal{C}_1'$ that contains all "compatible" level-1 codewords, i.e., all codewords encoding the tuple $(w, W_{d_2}^{(\mathrm{B})}, \ldots, W_{d_K}^{(\mathrm{B})})$ for some $w \in \{1, \ldots, \lfloor 2^{nR^{(\mathrm{A})}} \rfloor\}$. It then decodes its desired message $W_{d_1}^{(\mathrm{A})}$ using an optimal decoding rule for $\mathcal{C}_1'$. Notice that the rate of $\mathcal{C}_1'$ is the rate of $W_{d_1}^{(\mathrm{A})}$, and receiver 1's decoding performance is not degraded by the fact that the additional messages $W_{d_2}^{(\mathrm{B})}, \ldots, W_{d_K}^{(\mathrm{B})}$ are sent in level 1.

VI. Generalized Coded Caching

We only sketch the scheme for $K = 2$ and $t = 1$. See [23] for the general scheme and its analysis.

Split $W_d = \left(W_d^{(\mathrm{A})}, W_d^{(\mathrm{B})}\right)$, where $R^{(A)} = I(X; Y_1) - \epsilon$ and $R^{(\mathrm{B})} = I(X; Y_2) - \epsilon$ for some input distribution P_X and a small $\epsilon > 0$. Cache messages $W_1^{(\mathrm{B})}, \ldots, W_D^{(\mathrm{B})}$ at receiver 1 and messages $W_1^{(\mathrm{A})}, \ldots, W_D^{(\mathrm{A})}$ at receiver 2.

The transmitter zero pads the binary representation of $W_{d_1}^{(\mathrm{A})}$ to $nR^{(\mathrm{B})}$ bits and sends its XOR with $W_{d_2}^{(\mathrm{B})}$ over the channel using a point-to-point code $\mathcal{C}$. Receiver 2 decodes the XOR-message directly from $\mathcal{C}$, and XORs it with the zero-padded version of $W_{d_1}^{(\mathrm{A})}$ stored in its cache memory. Receiver 1 in contrast performs joint cache-channel decoding: It retrieves $W_{d_2}^{(\mathrm{B})}$ from its cache memory, and extracts a subcodebook $\mathcal{C}' \subseteq \mathcal{C}$ containing all codewords that are "compatible" with $W_{d_2}^{(\mathrm{B})}$. It then decodes its desired message $W_{d_1}^{(\mathrm{A})}$ using an optimal decoding rule for subcodebook $\mathcal{C}'$. Notice that the rate of $\mathcal{C}'$ is the rate of $W_{d_1}^{(\mathrm{A})}$, i.e., $I(X; Y_1) - \epsilon$.

VII. Summary and Conclusion

We provided close upper and lower bounds on the global capacity-memory tradeoff $\mathsf{C}^\star(\mathsf{M})$ of degraded BCs. The bounds coincide in the regimes of small and large total cache memory, characterized in terms of the BC statistics. Given a small total cache memory, it is optimal to assign it to the weakest receiver. In this regime, $\mathsf{C}^\star(\mathsf{M})$ grows as $\frac{\mathsf{M}}{D}$ which corresponds to a perfect global caching gain where all receivers can benefit from all the cache contents of the network. In the regime of moderate M, the slope of $\mathsf{C}^\star(\mathsf{M})$ seems to decrease as M increases. In this regime, we propose to allocate cache memories to all the receivers, but such that weaker receivers are provided with larger cache memories (as specified in Theorem 5). Once the total cache memory budget exceeds a certain threshold, it is optimal to uniformly allocate all the remaining cache memory across all the receivers and to store the same content in the extra portions of the receivers' cache memories. Here, $\mathsf{C}^\star(\mathsf{M})$ grows as $\frac{1}{K} \cdot \frac{\mathsf{M}}{D}$, which corresponds to a local caching gain.

We conclude that assigning the total cache memory uniformly across all the receivers is highly suboptimal over noisy broadcast channels in contrast to the noiseless setup considered in [1].

References

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

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

[3] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *arXiv*, 1605.02317, May, 2016.

[4] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Erasure broadcast networks with receiver caching," in Proc. *IEEE ISIT 2016*, Barcelona, Spain, July 10–15, 2016. pp. 1819–1823.

[5] 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.

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

[7] S. Wang, X. Tian and H. Liu, "Exploiting the unexploited of coded caching for wireless content distribution," in *Proc. IEEE ICNC, 2015.*

[8] 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.

[9] A. Ghorbel, M. Kobayashi, and S. Yang "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Trans. on Inf. Theory*, vol. 62, no. 11, pp. 6407–6422, Nov. 2016.

[10] S. Yang, K.-H. Ngo, and M. Kobayashi, "Content delivery with coded caching and massive MIMO in 5G," in Proc. of *9th ISTC*, Sep. 2016, pp. 370–374.

[11] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *arXiv*, 1511.0396.

[12] J. Hachem, N. Karamchandani, and S. Diggavi, "Content caching and delivery over heterogeneous wireless networks," in Proc. *2015 IEEE INFOCOM*, Hong Kong, China, Apr. 16–21, 2015, pp. 756–764.

[13] J. Zhang and P. Elia, "Feedback-aided coded caching for the MISO BC with small caches," *arXiv* 1606.05396.

[14] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "An upper bound on the capacity-memory tradeoff of begraded broadcast channels," in Proc. of *IEEE ISTC*, Brest, Sep., 2016.

[15] X. Yi and G. Caire, "Topological coded caching," in Proc. *ISIT*, July 2016, pp. 2039–2043.

[16] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Edge-facilitated wireless distributed computing," in Proc. *IEEE GLOBECOM*, Dec. 2016.

[17] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Intl. Symp. Inform. Theory,* Hong Kong China, June 2015.

[18] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv*, 1602.04207, February 2016.

[19] S. Pooya Shariatpanah, S. Abolfazl Motahari, B. Hossein Khalaj, "Multi-server coded caching," *IEEE Trans. on Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

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

[21] R. G. Gallager, "Capacity and coding for degraded broadcast channels," *Problem. Inf. Transm.*, vol 10, no. 3, pp. 3–14, 1974.

[22] S. Saeedi Bidokhti, M. Wigger, A. Yener, "Gaussian broadcast channels with receiver cache assignment," *submitted to ICC 2017.*

[23] S. Saeedi Bidokhti, M. Wigger, A. Yener, "Benefits of cache assignment on degraded Broadcast Channels," available on *stanford.edu/∼saeedi/caching.pdf.*