On the Capacity Region of a Cache-Aided Gaussian Broadcast Channel with Multi-Layer Messages

Mohammad Mohammadi Amiri and Deniz Gündüz
Electrical and Electronic Engineering Department, Imperial College London, London SW7 2BT, U.K.
Email: {m.mohammadi-amiri15, d.gunduz}@imperial.ac.uk

Abstract*—A cache-aided K-user Gaussian broadcast channel (BC) is studied. The transmitter has a library of N files, from which each user requests one. The users are equipped with caches of different sizes, which are filled without the knowledge of the user requests in a *centralized* manner. Differently from the literature, it is assumed that each file can be delivered to different users at different rates, which may correspond to different quality representations of the underlying content, e.g., scalable coded video segments. Accordingly, instead of a single achievable rate, the system performance is characterized by a rate tuple, which corresponds to the vector of rates users' requests can be delivered at. The goal is to characterize the set of all achievable rate tuples for a given total cache capacity by designing joint cache and channel coding schemes together with cache allocation across users. Assuming that the users are ordered in increasing channel quality, each file is coded into K layers, and only the first k layers of the requested file are delivered to user k, $k = 1, \dots, K$. Three different coding schemes are proposed, which differ in the way they deliver the coded contents over the BC; in particular, *time-division*, *superposition*, and *dirty paper coding* schemes are studied. Corresponding achievable rate regions are characterized, and compared with a novel outer bound. To the best of our knowledge, this is the first work studying the delivery of files at different rates over a cache-aided noisy BC.

I. Introduction

In the *coded caching* framework introduced in [1], transmission is performed over two phases: in the *placement phase*, which takes place during off-peak hours, users fill their caches without knowing the particular demands. Once the demands are revealed, they are satisfied simultaneously over the *delivery phase*. Here, we consider a Gaussian broadcast channel (BC) from the server to the users during the delivery phase. Cache-aided Gaussian BC is studied in [2] with and without fading, and in [3], [4] focusing on the high SNR regime. A packet-erasure BC is considered in [5] and [6]. A degraded BC is considered in [7], where the placement phase is performed in a centralized manner with the full knowledge of the channel during the delivery phase. In [8] delivery over a Gaussian BC is studied from an energy efficiency perspective, assuming that the channel conditions in the delivery phase are not known during the placement phase, for both centralized and decentralized caching scenarios.

In most of the existing literature on coded caching, the key assumption is that the files in the library are coded at a single common rate, and each user requests one file from the library in its entirety. Accordingly, the objective function in [5]–[7] is to maximize the common rate of the messages that can be delivered to all the users, and the supremum of the achievable rates is defined as the *capacity* of the caching network. In [9], the authors relaxed this assumption and allowed each user to request the files at a different quality, and equivalently, at a different rate. However, the required rates at which the contents must be delivered are assumed to be given as part of the problem definition in [9], and the goal is to find the minimum number of bits that must be delivered over an error-free shared delivery channel [9]. In this work, similarly to [9], we allow the users to request the files at different rates; however, differently from [9], considering a Gaussian BC in the delivery phase, we aim at characterizing the rate tuples at which the requested contents can be delivered to the users.

We argue that this formulation allows us to better exploit the asymmetric resources available to users for content delivery over a noisy BC. To see the difference between the scalar capacity definition used in [7] and the capacity region formulation proposed here, consider a Gaussian BC without any caches, i.e., $M = 0$. In this case, the capacity as defined in [7] is limited by the rate that can be delivered to the worst user, whereas with our formulation any rate tuple within the capacity region of the underlying BC is achievable, providing a much richer characterization of the performance for cache-aided delivery over a noisy BC.

The motivation here is to deliver the contents at higher rates to users with better channels, rather than being limited by the weak users. As proposed in [9], the multiple rates of the same file may correspond to the video files in the library encoded into multiple quality layers using scalable coding, so the user with a higher delivery rate receives a better quality description of the same file. Accordingly, each file in the library is coded into K layers, K being the number of users, ordered in increasing channel qualities, where user k receives layers 1 to k of its request, $k = 1, \dots, K$. We consider a centralized placement phase, and assume that the channel qualities of the users in the delivery phase are known in advance. Moreover, following [7], we consider a total cache capacity in the network as a constraint, and optimize cache allocation across the users and different layers of the files. Contents cached during the placement phase provide multicasting opportunities to the server to deliver the missing parts in the same layer of the files to different users. When delivering these coded contents to users over the underlying

This work was partially funded by the European Research Council through Starting Grant BEACON (agreement No. 677854).

BC, we consider three different techniques. Corresponding coding schemes are called joint cache and time-division coding (CTDC), joint cache and superposition coding (CSC), and joint cache and dirty paper coding (CDPC). We also present an outer bound on the rate region when the placement phase is constrained to uncoded caching, and compare it with the achievable rate tuples obtained though the proposed coding schemes.

Notations: For any arbitrary non-empty set $\mathcal{G}$ with cardinality $|\mathcal{G}|$, we denote the $\binom{|\mathcal{G}|}{i}$ i-element subsets of $\mathcal{G}$ by $\mathcal{S}^i_{\mathcal{G},1}, \dots, \mathcal{S}^i_{\mathcal{G},\binom{|\mathcal{G}|}{i}}$, for $i = 1, ..., |\mathcal{G}|$. For $g \notin \mathcal{G}$, we define $\{\mathcal{G}, g\} \triangleq \mathcal{G} \bigcup \{g\}$, and for $\mathcal{H} \subset \mathcal{G}$, $\mathcal{G} \backslash \mathcal{H}$ represents $\{j : j \in \mathcal{G}, j \notin \mathcal{H}\}$. For two integers i and j, $j \geq i$, $[i : j]$ denotes the set $\{i, i+1, ..., j\}$. For any positive real number q, we define $[q] \triangleq \{1, \dots, \lceil q \rceil\}$. We define, for two real values $p \geq 0$ and $q > 0$, $C^p_q \triangleq \frac{1}{2}\log_2(1 + p/q)$, and $\bar{p} \triangleq 1-p$. Notation $\bar{\oplus}$ represents bitwise XOR operation where the arguments are first zero-padded to have the same length as the longest argument.

II. System Model and Preliminaries

We consider cache-aided content delivery over a K-user Gaussian BC. The transmitter has a library of N files, $\mathbf{W} \triangleq W_1, ..., W_N$. File W_j is coded into K layers $W_j^{(1)}, \dots, W_j^{(K)}$, such that layer $W_j^{(l)}$ is distributed uniformly over the set $\left[2^{nR^{(l)}}\right]$, where $R^{(l)}$ represents the rate of the l-th layer and n denotes the blocklength, for $j = 1, ..., N$, and $l = 1, ..., K$. We denote the i-th layers of all the files by $\mathbf{W}^{(i)} \triangleq W_1^{(i)}, ..., W_N^{(i)}$, for $i \in [K]$.

Assume that user k, $k \in [K]$, has a cache of capacity nM_k bits, which is filled during the *placement phase* without the knowledge of the user demands. User demands are revealed and satisfied simultaneously in the *delivery phase*. Each user requests a single file from the library, where W_{d_k}, $d_k \in [N]$, denotes the file requested by user $k \in [K]$. For a demand vector $\mathbf{d} \triangleq (d_1, ..., d_K)$, the users are served by a common message $X^n(\mathbf{W}) \triangleq (X_1(\mathbf{W}), \dots, X_n(\mathbf{W}))$ satisfying the average power constraint. User k, $k \in [K]$, receives $Y_k^n(\mathbf{W}) \triangleq (Y_{k,1}(\mathbf{W}), \dots, Y_{k,n}(\mathbf{W}))$ through a Gaussian channel

$$Y_k^n(\mathbf{W}) = X^n(\mathbf{W}) + Z_k^n, \tag{1}$$

where $Z_k^n \triangleq (Z_{k,1}, \dots, Z_{k,n})$, and $Z_{k,i}$ is the independent zero-mean real Gaussian noise with variance σ_k^2 at user k at the i-th channel use. Without loss of generality we order the users in increasing channel quality, i.e., we assume that $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_K^2$. We define $\boldsymbol{\sigma} \triangleq (\sigma_1, \dots, \sigma_K)$.

Placement phase is performed in a centralized manner assuming $\boldsymbol{\sigma}$ is known. An $\left(n, R^{(1)}, \dots, R^{(K)}, M_1, \dots, M_K\right)$ code consists of the following:

- K caching functions ϕ_k, $k \in [K]$, where ϕ_k maps $\mathbf{W}$ and $\boldsymbol{\sigma}$ to the cache content U_k of user k, i.e., $U_k = \phi_k(\mathbf{W}, \boldsymbol{\sigma})$.
- An encoding function ψ, which generates the channel input as $X^n(\mathbf{W}) = \psi(\mathbf{W}, \boldsymbol{\sigma}, \mathbf{d})$, for demand vector $\mathbf{d}$, satisfying the average power constraint $\frac{1}{n}\sum_{i=1}^{n} X_i^2(\mathbf{W}) \leq P$.
- K decoding functions $\mu_{\mathbf{d},k}$, $k \in [K]$, where, for a demand vector $\mathbf{d}$, $\mu_{\mathbf{d},k}$ reconstructs the layers $\hat{W}_{d_k}^{(1)}, \dots, \hat{W}_{d_k}^{(k)}$ from the channel output Y_k^n and cache content U_k.

The probability of error is defined as $P_e \triangleq \Pr\left\{ \bigcup_{\mathbf{d} \in [N]^K} \bigcup_{k=1}^{K} \bigcup_{i=1}^{k} \left\{ \hat{W}_{d_k}^{(i)} \neq W_{d_k}^{(i)} \right\} \right\}$.

Note that the generated code implicitly assumes that user k is interested only in the first k layers of its demand, i.e., $W_{d_k}^{(1)}, \dots, W_{d_k}^{(k)}$, for $k \in [K]$. Therefore, its cache contents depend only on $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)}$. In a more general formulation, we could instead consider an arbitrary ordering of the rates among the users, but here the goal is to deliver a higher rate to a user with a better channel.

For a given total cache capacity M, we say that the rate tuple $(R_1, \dots, R_K)$ is achievable if for every $\varepsilon > 0$, there exists an $\left(n, R^{(1)}, \dots, R^{(K)}, M_1, \dots, M_K\right)$ code, which satisfies $P_e < \varepsilon$, $R_k \leq \sum_{l=1}^{k} R^{(l)}$, and $\sum_{k=1}^{K} M_k \leq M$. For average power constraint P and a total cache capacity M, the capacity region $\mathcal{C}(P, M)$ of the caching system described above is defined as the closure of the all achievable rate tuples. Our goal is to find inner and outer bounds on $\mathcal{C}(P, M)$.

Next, we present some definitions that will simplify our ensuing presentation. For a fixed value of t, $t \in [K-1]$, we define $g_l \triangleq \sum_{j=1}^{l} \binom{K-j}{t}$, $\forall l \in [K-t]$, and let $g_0 = 0$. We note that $g_{K-t} = \binom{K}{t+1}$. We denote the set of users $[l : K]$ by $\mathcal{K}_l$, for $l \in [K]$. We label $(t+1)$-element subsets of users in $\mathcal{K}_1$, so that the subsets with the smallest element l are labelled as

$$\mathcal{S}^{t+1}_{\mathcal{K}_1, 1+g_{l-1}}, \dots, \mathcal{S}^{t+1}_{\mathcal{K}_1, g_l}, \quad \text{for } l = 1, ..., K-t. \tag{2}$$

Thus, we have, for $l \in [K-t]$,

$$\left\{\mathcal{S}^{t+1}_{\mathcal{K}_1, 1+g_{l-1}} \backslash \{l\}, \dots, \mathcal{S}^{t+1}_{\mathcal{K}_1, g_l} \backslash \{l\}\right\} = \left\{\mathcal{S}^{t}_{\mathcal{K}_{l+1}, 1}, \dots, \mathcal{S}^{t}_{\mathcal{K}_{l+1}, \binom{K-l}{t}}\right\}, \tag{3}$$

i.e., the family of all $(t+1)$-element subsets of $\mathcal{K}_1$ excluding l, which is their smallest element, is the same as the family of all t-element subsets of $\mathcal{K}_{l+1}$. We note that the number of subsets of users in both sets in (3) is $\binom{K-l}{t}$, $l \in [K-t]$. Without loss of generality, we label the subsets of users so that, for $l \in [K-t]$,

$$\mathcal{S}^{t+1}_{\mathcal{K}_1, i+g_{l-1}} \backslash \{l\} = \mathcal{S}^{t}_{\mathcal{K}_{l+1}, i}, \text{ for } i \in \left[\binom{K-l}{t}\right]. \tag{4}$$

III. Achievable Schemes

Here we present three different inner bounds on $\mathcal{C}(P, M)$.

A. Joint Cache and Time-Division Coding (CTDC)

With CTDC, the missing bits corresponding to the layers in $\mathbf{W}^{(l)}$ are delivered in a coded manner exploiting the cached contents as in the standard coded caching framework. The coded contents are transmitted over the BC using time-division among layers. We elaborate on the placement and delivery phases of the CTDC scheme in the longer version of the paper [10].

Proposition 1. *For the system described in Section II with average power P and total cache capacity M, the rate tuple $(R_1, ..., R_K)$ is achievable by the CTDC scheme, if there exist $t_1, \dots, t_K$, where $t_l \in [0 : K-l]$, $\forall l \in [K]$, non-negative $R^{(1)}, \dots, R^{(K)}$, and non-negative $\lambda^{(1)}, \dots, \lambda^{(K)}$, such that $R_k = \sum_{l=1}^{k} R^{(l)}$, $\sum_{l=1}^{K} \lambda^{(l)} = 1$, $\forall k \in [K]$, and*

$$R^{(l)} \le \lambda^{(l)} \frac{\sum_{i=1}^{\binom{K-l+1}{t_l}} \prod_{k \in \mathcal{K}_l \backslash \mathcal{S}_{\mathcal{K}_l,i}^{t_l}} C_{\sigma_k^2}^{P}}{\sum_{i=1}^{\binom{K-l+1}{t_l+1}} \prod_{k \in \mathcal{K}_l \backslash \mathcal{S}_{\mathcal{K}_l,i}^{t_l+1}} C_{\sigma_k^2}^{P}}, \quad \text{for } l \in [K], \tag{5a}$$

$$M = N \sum_{l=1}^{K} t_l R^{(l)}. \tag{5b}$$

Corollary 1. *The following rate region for a total cache capacity M and average power P can be achieved by the CTDC scheme:*

$$\mathcal{C}_b(P, M) = \bigcup_{\lambda^{(1)}, \dots, \lambda^{(K)} : \sum_{l=1}^{K} \lambda^{(l)} = 1} \left(\{R_1, \dots, R_K\} : (R_1, \dots, R_K) \text{ and } M \text{ satisfy } (5) \right). \tag{6}$$

Remark 1. *Let $(\hat{R}_1, \dots, \hat{R}_K) \in \mathcal{C}_b(P, M)$ and $(\tilde{R}_1, \dots, \tilde{R}_K) \in \mathcal{C}_b(P, M)$. Then, for any $\lambda \in [0, 1]$, $(\lambda \hat{R}_1 + \bar{\lambda} \tilde{R}_1, \dots, \lambda \hat{R}_K + \bar{\lambda} \tilde{R}_K) \in \mathcal{C}_b(P, M)$. This can be shown by joint time and memory-sharing. The whole library is divided into two parts according to λ, and the delivery of the two parts are carried out over two orthogonal time intervals of length λn and $\bar{\lambda} n$ using the codes for the two achievable tuples. Thus, for a fixed total cache capacity M, the rate pairs in the convex-hull of $\mathcal{C}_b(P, M)$ are achievable.*

According to Remark 1, a rate vector $\mathbf{R}^* \triangleq (R_1^*, \dots, R_K^*)$ is on the boundary surface of $\mathcal{C}_b(P, M)$, if there exist non-negative coefficients $w_1, \dots, w_K$, $\sum_{i=1}^{K} w_i = 1$, for which $\mathbf{R}^*$ is a solution to the following optimization problem:

$$\max_{\lambda^{(1)}, \dots, \lambda^{(K)}, R_1, \dots, R_K} \sum_{i=1}^{K} w_i R_i, \quad \text{subject to } \{R_1, ..., R_K\} \in \mathcal{C}_b(P, M). \tag{7}$$

In the other words, for given weights $w_1, \dots, w_K$, and total cache capacity M, $\mathbf{R}^*$ solves the problem in (7), if $R^{(1)}, \dots, R^{(K)}$ is a solution of the following problem:

$$\max_{\lambda^{(1)}, \dots, \lambda^{(K)}, R^{(1)}, \dots, R^{(K)}} \sum_{i=1}^{K} w_i \sum_{l=1}^{i} R^{(l)}, \quad \text{subject to (5a) and (5b)}, \quad \sum_{l=1}^{K} \lambda^{(l)} = 1, \tag{8}$$

and

$$R_k^* = \sum_{l=1}^{k} R^{(l)}, \quad \text{for } k = 1, \dots, K. \tag{9}$$

B. Joint Cache and Superposition Coding (CSC) and Joint Cache and Dirty Paper Coding (CDPC)

Here we present the achievable rate regions for the CSC and CDPC schemes. We introduce r_1 and r_2 to distinguish between the two, where we set $r_1 = 0$ and $r_2 = 1$ for CSC, while $r_1 = 1$ and $r_2 = 0$ for CDPC. We briefly highlight here that, with the CSC scheme, the coded packets of different layers are delivered over the Gaussian BC through superposition coding, while the CDPC scheme uses dirty paper coding to deliver the coded packets of different layers. The CSC scheme along with an example highlighting the main techniques and the CDPC scheme are elaborated in the longer version of the paper [10].

Theorem 1. *For the system described in Section II with average power P and total cache capacity M, rate tuple $(R_1, ..., R_K)$ is achievable, if there exist $t \in [K-1]$, and non-negative $R^{(1)}, \dots, R^{(K)}$, such that $R_k = \sum_{l=1}^{k} R^{(l)}$, for $k \in [K]$, and*

$$R^{(l)} = \begin{cases} \sum_{i=1}^{\binom{K}{t}} R^{(1)}_{\mathcal{S}^t_{\mathcal{K}_1,i}}, & \text{if } l = 1, \\ \sum_{i=1}^{\binom{K-l+1}{t-1}} R^{(l)}_{\mathcal{S}^{t-1}_{\mathcal{K}_l,i}}, & \text{if } l = 2, ..., K-t+1, \\ 0, & \text{otherwise}, \end{cases} \tag{10a}$$

and, for $i \in [1 + g_{l-1} : g_l]$ and $l \in [K-t]$,

$$R^{(1)}_{\mathcal{S}^{t+1}_{\mathcal{K}_1,i} \backslash \{k_1\}} \le \lambda_i C^{\alpha_i P}_{\bar{\alpha}_i P r_2 + \sigma^2_{k_1}}, \ \forall k_1 \in \mathcal{S}^{t+1}_{\mathcal{K}_1,i}, \tag{10b}$$

$$R^{(l+1)}_{\mathcal{S}^t_{\mathcal{K}_{l+1},i-g_{l-1}} \backslash \{k_2\}} \le \lambda_i C^{\bar{\alpha}_i P}_{\alpha_i P r_1 + \sigma^2_{k_2}}, \ \forall k_2 \in \mathcal{S}^t_{\mathcal{K}_{l+1},i-g_{l-1}}, \tag{10c}$$

and

$$M = N \left(t R^{(1)} + (t-1) \sum_{l=2}^{K-t+1} R^{(l)} \right), \tag{10d}$$

for some

$$0 \le \alpha_i \le 1, \tag{10e}$$

$$0 \le \lambda_i \le 1, \quad \text{for } i = 1, ..., \tbinom{K}{t+1}, \tag{10f}$$

$$\sum_{i=1}^{\binom{K}{t+1}} \lambda_i = 1. \tag{10g}$$

Corollary 2. *The following rate region for a total cache capacity M and average power constraint P can be achieved:*

$$\mathcal{C}_c(P, M) = \bigcup_{\boldsymbol{\alpha}, \boldsymbol{\lambda} : \sum_{i=1}^{\binom{K}{t+1}} \lambda_i = 1} \left(\{R_1, \dots, R_K\} : (R_1, \dots, R_K) \text{ and } M \text{ satisfy } (10) \right), \tag{11}$$

where $\boldsymbol{\alpha} \triangleq \alpha_1, \dots, \alpha_{\binom{K}{t+1}}$, and $\boldsymbol{\lambda} \triangleq \lambda_1, \dots, \lambda_{\binom{K}{t+1}}$.

For a fixed total cache capacity M, the convexity of region $\mathcal{C}_c(P, M)$ is followed through the same argument as Remark 1, for both the CSC and CDPC schemes. As a result, for a given total cache capacity M, and for given non-negative coefficients $w_1, \dots, w_K$, such that $\sum_{i=1}^{K} w_i = 1$, a rate vector $\mathbf{R}^*$ is on the boundary surface of the achievable rate region $\mathcal{C}_c(P, M)$, if $R^{(1)}, \ldots, R^{(K)}$ is a solution of the following problem:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\lambda},\mathbf{R}^{(1)},\ldots,\mathbf{R}^{(K-t+1)}} \sum_{i=1}^{K} w_i \sum_{l=1}^{i} R^{(l)},$$
$$\text{subject to } R^{(1)}, \ldots, R^{(K-t+1)} \text{ satisfy (10a)},$$
$$\mathbf{R}^{(1)} \text{ satisfy (10b)},$$
$$\mathbf{R}^{(2)}, \ldots, \mathbf{R}^{(K-t+1)} \text{ satisfy (10c)},$$
$$M \text{ satisfies (10d)},$$
$$\boldsymbol{\alpha} \text{ and } \boldsymbol{\lambda} \text{ satisfy (10e)-(10g)}, \quad (12a)$$

where

$$\mathbf{R}^{(1)} \triangleq R^{(1)}_{\mathcal{S}^t_{\mathcal{K}_1,1}}, \ldots, R^{(1)}_{\mathcal{S}^t_{\mathcal{K}_1,\binom{K}{t}}}, \quad (12b)$$

$$\mathbf{R}^{(l)} \triangleq R^{(l)}_{\mathcal{S}^t_{\mathcal{K}_l,1}}, \ldots, R^{(l)}_{\mathcal{S}^t_{\mathcal{K}_l,\binom{K-l+1}{t-1}}}, \text{ for } l \in [2 : K-t+1], \quad (12c)$$

and

$$R^*_k = \sum_{l=1}^{k} R^{(l)}, \quad \text{for } k = 1, \ldots, K. \quad (13)$$

Remark 2. *Let $\tilde{\mathbf{R}} \triangleq (\tilde{R}_1, \ldots, \tilde{R}_K)$ and $\hat{\mathbf{R}} \triangleq (\hat{R}_1, \ldots, \hat{R}_K)$ be two achievable rate tuples for total cache capacities $\tilde{M}$ and $\hat{M}$, respectively. Then, $\beta\tilde{\mathbf{R}} + \bar{\beta}\hat{\mathbf{R}}$ can be achieved through joint time and memory-sharing for a total cache capacity $\beta\tilde{M} + \bar{\beta}\hat{M}$, for some $\beta \in [0,1]$. For $M = 0$, the system under consideration is equivalent to the Gaussian BC without user caches, where user k requests a file of rate $\sum_{l=1}^{k} R^{(l)}$, $k \in [K]$, and rate tuple $\mathbf{R}_z \triangleq (R_{z_1}, ..., R_{z_K})$ is achievable by superposition coding, where*

$$R_{z_k} = C^{\gamma_k P}_{\sum_{i=k+1}^{K} \gamma_i P + \sigma_k^2}, \quad \text{for } k = 1, ..., K, \quad (14)$$

for some non-negative coefficients $\gamma_1, \ldots, \gamma_K$, such that $\sum_{i=1}^{K} \gamma_i = 1$. Hence, rate tuples $\beta\mathbf{R}_z + \bar{\beta}\tilde{\mathbf{R}}$ and $\beta\mathbf{R}_z + \bar{\beta}\hat{\mathbf{R}}$ are also achievable for total cache capacities $\bar{\beta}\tilde{M}$ and $\bar{\beta}\hat{M}$, respectively, through time sharing.

IV. Outer Bound

In the following, we present an outer bound on the capacity region $\mathcal{C}(P, M)$, whose proof can be found in the longer version of the paper [10, Appendix C].

Theorem 2. *Consider the system described in Section II with average power P, where user k has a cache capacity of M_k, $k \in [K]$. If the placement phase is constrained to uncoded caching, for any non-empty subset $\mathcal{G} \subset [K]$, we have, for $k = 1, \ldots, |\mathcal{G}|$,*

$$R_{\pi_{\mathcal{G}}(k)} \leq C^{\eta_{\pi_{\mathcal{G}}(k)} P}_{\sum_{i=k+1}^{|\mathcal{G}|} \eta_{\pi_{\mathcal{G}}(i)} P + \sigma^2_{\pi_{\mathcal{G}}(k)}} + \frac{1}{N} \sum_{i=1}^{k} M_{\pi_{\mathcal{G}}(i)}, \quad (15)$$

for some non-negative coefficients $\eta_{\pi_{\mathcal{G}}(1)}, \ldots, \eta_{\pi_{\mathcal{G}}(|\mathcal{G}|)}$, such that $\sum_{i=1}^{|\mathcal{G}|} \eta_{\pi_{\mathcal{G}}(i)} = 1$, where $\pi_{\mathcal{G}}$ is a permutation of the elements of $\mathcal{G}$, such that $\sigma^2_{\pi_{\mathcal{G}}(1)} \geq \sigma^2_{\pi_{\mathcal{G}}(2)} \geq \cdots \geq \sigma^2_{\pi_{\mathcal{G}}(|\mathcal{G}|)}$.

Remark 3. *The outer bound is not tight in general, particularly when the channel qualities are more skewed. This is due to the nature of the underlying model, where cache allocation is allowed, and the capacity is characterized as a function of the available total cache capacity, whereas the outer bound is specified for a particular cache allocation. Moreover, unlike the model studied in [7], the asymmetry due to different rate delivery to different users increases the gap between the outer bound and the achievable schemes.*

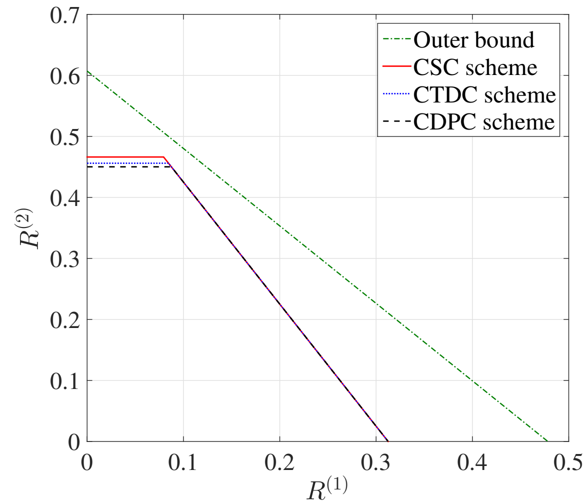

Fig. 1. Achievable rate pair $(R^{(1)}, R^{(2)})$ for a caching system with $K = N = 4$, and $M = 2.5$, where $R^{(3)} = 0$, $t = 2$ and $t_1 = 2$, $t_2 = t_3 = 1$, and $t_4 = 0$. The noise variance at user k is $\sigma_k^2 = 5 - k$, for $k = 1, ..., 4$, and we set $P = 2$.

V. Numerical Results

In this section, we compare the achievable rate regions of the CTDC, CSC, and CDPC schemes for a caching system with $K = N = 4$. We set the average power constraint to $P = 2$, and the noise variance at user k is assumed to be $\sigma_k^2 = 5 - k$, for $k \in [4]$. We assume a total cache capacity of $M = 2.5$.

We evaluate the performance in terms of the rate of different layers of the files, i.e., $R^{(1)}, \ldots, R^{(K)}$, where $R_k = \sum_{l=1}^{k} R^{(l)}$, for $k \in [K]$. We examine the performance of the CSC and CDPC schemes for $t = 2$. Thus, the achievable rate tuple (R_1, R_2, R_3, R_4) presented in Theorem 1 can be achieved by the CSC and CDPC schemes, for $r_1 = 0$, $r_2 = 1$, and $r_1 = 1$, $r_2 = 0$, respectively, where $R_4 = R_3$ since $R^{(4)} = 0$. The boundary surface of the rate region achieved by the CSC and CDPC schemes are computed through the optimization problem given in (12). For the fairness of the comparison, we consider caching factors $t_1 = 2$, $t_2 = t_3 = 1$, and $t_4 = 0$. The boundary of the rate region achievable by CTDC can be calculated by the optimization problem in (8), where, in order to have a fair comparison, we set $\lambda^{(4)} = 0$ leading to $R^{(4)} = 0$ and $R_4 = R_3$.

We investigate the convex hull of the achievable rate tuples calculated by the optimization problem corresponding to each of the CTDC, CSC, and CDPC schemes. Since the presentation of the three-dimensional rate region together with the

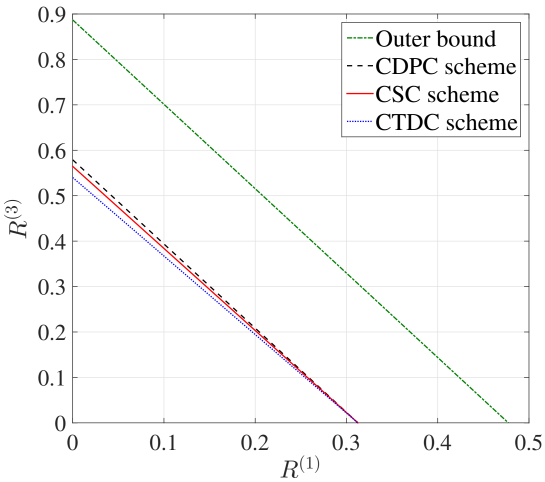

Fig. 2. Achievable rate pair $(R^{(1)}, R^{(3)})$ for a caching system with $K = N = 4$, and $M = 2.5$, where $R^{(2)} = 0$, and $t = 2$ and $t_1 = 2, t_2 = t_3 = 1$, and $t_4 = 0$. The noise variance at user k is $\sigma_k^2 = 5 - k$, for $k = 1, ..., 4$, and we set $P = 2$.

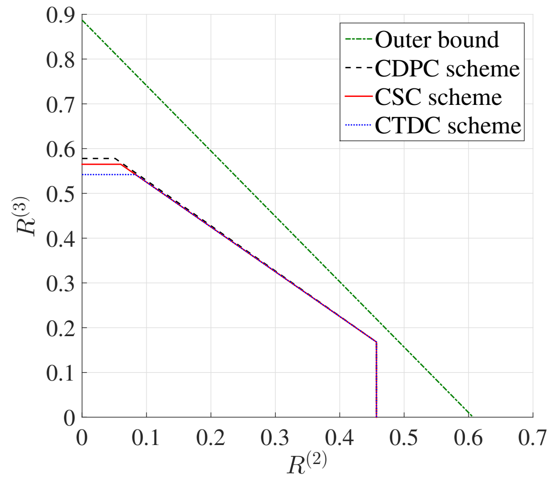

Fig. 3. Achievable rate pair $(R^{(2)}, R^{(3)})$ for a caching system with $K = N = 4$, and $M = 2.5$, where $R^{(1)} = 0$, and $t = 2$ and $t_1 = 2, t_2 = t_3 = 1$, and $t_4 = 0$. The noise variance at user k is $\sigma_k^2 = 5 - k$, for $k = 1, ..., 4$, and we set $P = 2$.

outer bound does not provide a clear picture, here we fix one of the rates $R^{(1)}$, $R^{(2)}$ and $R^{(3)}$ and present the rate region on the two-dimensional planes corresponding to the other two rates. Two-dimensional plane of $(R^{(1)}, R^{(2)})$, $(R^{(1)}, R^{(3)})$ and $(R^{(2)}, R^{(3)})$ for $R^{(3)} = 0$, $R^{(2)} = 0$ and $R^{(1)} = 0$ are illustrated in in Figures 1, 2 and 3, respectively, together with the outer bound presented in Theorem 2. As it can be seen from the figures, for relatively small values of $R^{(1)}$, the CSC and CTDC schemes achieve higher values of $R^{(2)}$, while the CSC scheme outperforms the latter. For higher values of $R^{(1)}$, the improvement of the CSC scheme over CTDC and CDPC is negligible. For a fixed $R^{(1)}$ value, CDPC achieves higher values of $R^{(3)}$ compared to the other two achievable schemes, and CSC outperforms CTDC. On the other hand, given a relatively small value of $R^{(2)}$, CDPC improves upon the CSC and CTDC in terms of the achievable rate $R^{(3)}$, and CSC achieves higher values of $R^{(3)}$ than CTDC. As mentioned in Remark 3, the outer bound is not tight in general; however, for any achievable rate tuple $(R_1, \dots, R_4)$, which is achieved with a specific cache allocation $M_1, \dots, M_4$, the outer bound specialized to this cache allocation would be tighter.

VI. Conclusions

We have studied cache-aided content delivery over a Gaussian BC, where each user is allowed to demand a file at a distinct rate. To model this asymmetry, we have assumed that the files are encoded into K layers corresponding to K users in the system, such that the k-th worst user is delivered only the k layers of its demand, $k \in [K]$. We have considered a centralized placement phase, where the server knows the channel qualities of the links in the delivery phase in addition to the identity of the users. By allowing the users to have different cache capacities, we have defined the capacity region for a total cache capacity. We designed a placement phase through cache allocation across the users and the files' layers to maximize the rates allocated to different layers. We have proposed three achievable schemes, which deliver coded multicast packets, generated thanks to the contents carefully cached during the placement phase, through different channel coding techniques over the Gaussian BC. Although the coded multicast packets are intended for a set of users with distinct link capacities, channel coding techniques can be employed to deliver requested files such that the users with better channels achieve higher rates. We have also developed an outer bound on the capacity region assuming uncoded caching. We are currently working to reduce the gap between the inner and outer bounds.

References

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

[2] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv:1504.01452v1 [cs.IT]*, Apr. 2015.

[3] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *arXiv:1711.05969 [cs.IT]*, Nov. 2017.

[4] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," in *Proc. IEEE Int'l Symp. on Inform. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 401–405.

[5] S. Saeedi Bidokhti, R. Timo, and M. Wigger, "Noisy broadcast networks with receiver caching," *arXiv:1605.02317v1 [cs.IT]*, May 2016.

[6] M. Mohammadi Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

[7] S. Saeedi Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *arXiv:1702.08044v1 [cs.IT]*, Feb. 2017.

[8] M. Mohammadi Amiri and D. Gündüz, "Caching and coded delivery over Gaussian broadcast channels for energy efficiency," *arXiv:1712.03433v2 [cs.IT]*, *IEEE J. Sel. Areas Commun.*, to appear.

[9] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," *IEEE Trans. Inform. Theory*, Feb. 2018.

[10] M. Mohammadi Amiri and D. Gündüz, "On the capacity region of a cache-aided Gaussian broadcast channel with multi-layer messages," in preparation.