Vector Coded Caching Substantially Boosts MU-MIMO: Pathloss, CSI and Power-allocation Considerations

Hui Zhao, and Petros Elia
Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France
Email: hui.zhao@eurecom.fr; petros.elia@eurecom.fr

***Abstract*—We here continue the exploration of vector coded caching and its powerful effect on the performance of multi-user multiple-input multiple-output (MU-MIMO) systems. In particular, this paper investigates the impact that vector coded caching under various practical considerations, which include channel state information acquisition costs, multi-antenna receivers, variable pathloss, as well as max-min fairness where the minimum transmission rate among the simultaneously served users is maximized via power allocation. We here focus on the widely adopted zero-forcing precoder, and proceed to show the large multiplicative boost brought about by vector coded caching over the optimized cacheless counterpart. To do so, we here provide lower and upper bounds on the overall throughput as well as on the effective gain (i.e., on the rate improvement over the optimized cacheless counterpart), again in the presence of the aforementioned practical aspects. Finally various numerical results demonstrate the tightness of the derived upper and lower bounds, as well as nicely illustrate the notable gains.**

I. Introduction

The seminal work in [1] revealed that using caches at the receivers could allow a transmitter to simultaneously serve multiple users over a rank-one error-free shared-link broadcast channel with K cache-aided receiving users requesting different content from a content library of files. The new method, widely known as coded caching, was able to exploit the ability of each user to store a fraction $\gamma \in [0,1]$ of the content library, in order to yield $K\gamma + 1$ degrees of freedom (DoF) which matches what is also known as the *theoretical caching gain*.

Since then, there has been considerable effort to explore the practical ramifications of this astounding theoretic breakthrough. One of the first realizations is that unfortunately, the real gains of coded caching would need — under realistic considerations — to be substantially downgraded because of the combinatorial property of the algorithm in [1] that required each file to be split into an astronomical number of $\binom{K}{K\gamma}$ non-overlapping subfiles. In the presence of finite file sizes, the real DoF would then have to be reduced to a dramatically smaller $\Lambda\gamma + 1$, where $\Lambda \ll K$ is determined by the file size. As of now, under most practical constraints, this DoF of $\Lambda\gamma + 1$ remains clearly within the single-digit range [2], [3].

Such bottlenecks led to the realization that coded caching must be considered in conjunction with other resources, such as for example multi-antenna arrays which are abundantly used in various wireless networks. This brought to light the area of *multi-antenna coded caching* which considers coded caching as a means for boosting multi-user multiple-input multiple-output (MU-MIMO) systems. Related works in this new area include the original works of [4], [5], or the work in [6] which considered various physical-layer (PHY)-inspired multi-antenna coded caching schemes over symmetric Rayleigh fading channels. The scalability of content delivery rate for coded caching was investigated in massive MIMO and cell-free networks respectively in [7], [8]. More recently the work in [9] designed cache-aided multi-antenna precoders that appropriately account for inter-stream interference, while the work in [10] proposed a location-dependent multi-antenna coded caching scheme tailored for wireless extended reality applications. It is the case though that all the above schemes consider, one way or another, the use of multiple antennas as a means for delivering the multicast messages from [1], and for this reason, due to the aforementioned severity of the file-size constraint, their effective/real DoF is solidly bounded by $\Lambda\gamma + Q$, where Q is the chosen multiplexing gain provided by multiple transmit antennas. As a consequence, *the impact of caching had remained dwarfed by the existing and available multiplexing gains*, especially in the massive MIMO regime where $\Lambda\gamma \ll Q$, which has also been extensively demonstrated in field trails [11].

The impact of caching on multi-antenna systems was substantially increased with the introduction of vector coded caching in [3], where a clique-based structure on vectors is employed, in place of the aforementioned multicasting-based (XOR-based) transmissions that governed previous multi-antenna coded caching schemes. In the context of file-size constrained communications, vector coded caching provides a DoF of $Q(\Lambda\gamma + 1)$ which implies a multiplicative DoF boost over the cacheless case (whose DoF would be Q), where now the impact of caching far exceeds the additive impact (see DoF of $\Lambda\gamma + Q$) of previous XOR-based multi-antenna coded caching approaches. This performance also comes at a controllable subpacketization cost. Following these findings that considered the high-SNR (DoF) regime, the works in [12]–[14] further investigated the performance of vector coded caching in the presence now of finite and realistic SNRs regimes, single-antenna receivers and symmetric Rayleigh fading channels, by providing analysis that captures gains from beamforming and costs for gathering channel state information (CSI).

Apart from these preliminary works [12]–[14], the effect of many other practical aspects on vector coded caching, still remains unknown. Three important such ingredients which we consider here include pathloss, the use of multi-

This work is supported by the European Research Council under the EU Horizon 2020 research and innovation program/ERC grant agreement no. 725929 (ERC project DUALITY).

antenna receivers, and power-allocation. Regarding the first such ingredient, it is indeed the case that understanding the effect of pathloss importantly relates directly to the dreaded near-far bottleneck resulting from having users with different pathloss, where this bottleneck is widely recognized as one of the main challenges in applying coded caching in wireless settings [15]–[18]. Similarly the multi-antenna aspect is also important because now, more and more, various 5G and beyond-5G scenarios include not only multi-antenna base-stations (BSs), but also multi-antenna receiving users; an aspect that has sparked particular research interest [19], [20] in the context of coded caching. Last but not least, the performance of vector coded caching is naturally affected by the adopted power allocation algorithms in practical SNRs, especially as we need to satisfy various user fairness requirements [9], [21], such as for example max-min fairness (MMF), which are often designed to help some of the ill-positioned users.

We here explore the performance of vector coded caching in practical MU-MIMO scenarios, additionally incorporating these issues, thus considering — in addition to channel fading, CSI costs, and linear precoding — also multi-antenna receivers, pathloss, as well as MMF. The corresponding closed-form expressions for the lower and upper bounds of the overall throughput and of the effective gain are derived for zero-forcing (ZF) precoding, which significantly facilitates the performance evaluation due to avoiding complicated and time-consuming numerical optimization for MMF. Our numerical results validate the tightness of the derived bounds under realistic assumptions, as well as clearly illustrate the notable gains brought about by vector coded caching.

Notations: For a positive integer a, we use $[a]$ to denote the set $\{1, 2, \cdots, a\}$. $|\cdot|$ denotes either the cardinality for a set or the magnitude for a complex number. $\mathbb{C}$ denotes the set of complex numbers, while $\mathbb{E}\{\cdot\}$ and $\mathrm{Tr}\{\cdot\}$ represents the average operator and the trace operator respectively. $\mathrm{Diag}\{a_1, \cdots, a_n\}$ denotes a diagonal matrix with the diagonal elements $a_1, \cdots, a_n$, while $\Psi \setminus \phi$ denotes the set Ψ with removing the element ϕ. We use $\mathbf{A}^T$ and $\mathbf{A}^H$ to denote the non-conjugate transpose and conjugate transpose of the matrix $\mathbf{A}$ respectively. $[\mathbf{A}]_{\ell,\vartheta}$ denotes the (ℓ, ϑ)-th element in the matrix $\mathbf{A}$. $\mathbf{0}_L \in \mathbb{C}^L$ denotes the vector with all zero elements, and $\mathbf{I}_L \in \mathbb{C}^{L\times L}$ represents the identity matrix.

II. System Model and Problem Description

In a cache-aided downlink MU-MIMO system, a BS equipped with L antennas serves K cache-aided users where each user is equipped with M antennas and requests a different file from a library $\mathcal{F}$ with N $(N \geq K)$ equal-sized files in total. The BS has full access to the library $\mathcal{F}$ while each user can only cache a fraction $\gamma \in [0, 1]$ of library content during off-peak hours. To accelerate the delivery, we slightly modify the vector coded caching scheme originally proposed in [3] for single-antenna users, which we will detail in Section II-A. Due to the finite file-size constraint, there are only $\Lambda \leq K$ different cache states in vector coded caching, which (as elaborated in [3]) leads to having $B = \frac{K}{\Lambda}$ users sharing the same cache state, i.e., storing the same content in their individual caches during the cache-placement phase. Again as elaborated in [3], during the subsequent delivery phase of vector coded caching, there are $G \triangleq \Lambda\gamma + 1$ user groups (each group with their own cache state) that are simultaneously selected for service in each transmission round, and where we have $Q \in [B]$ users associated to each selected cache state that are active for the purposes of receiving messages during that instance. This implies that there are as many as GQ users served at a time.

We use Ψ to denote the selected G cache states for service, and use $\mathrm{U}_{\psi,k}$ to denote the k-th $(k \in [Q])$ active user in the cache state $\psi \in \Psi$. As a result of having multi-antenna receivers, the BS can send M symbols to $\mathrm{U}_{\psi,k}$ simultaneously, thereby enhancing the throughput to $\mathrm{U}_{\psi,k}$. We use $\mathbf{s}_{\psi,k} \triangleq [s_{\psi,k,1}, \cdots, s_{\psi,k,M}]^T \in \mathbb{C}^M$ and $\mathbf{P}_{\psi,k} \triangleq \mathrm{Diag}\{\sqrt{P_{\psi,k,1}}, \cdots, \sqrt{P_{\psi,k,M}}\} \in \mathbb{C}^{M\times M}$ to denote the data vector to $\mathrm{U}_{\psi,k}$ and the corresponding power allocation matrix respectively, where the independent variables $\{s_{\psi,k,q} : q \in [M]\}$ have zero-mean and unit-power. The signal streams $\mathbf{s}_{\psi,k}$ to $\mathrm{U}_{\psi,k}$ will be generally processed to $\mathbf{V}_{\psi,k}\mathbf{s}_{\psi,k}$ by a precoding matrix $\mathbf{V}_{\psi,k} \in \mathbb{C}^{L\times M}$. The transmitted signal $\mathbf{x}_\Psi \in \mathbb{C}^L$ for the selected user groups Ψ at the BS in the PHY is designed as[1]

$$\mathbf{x}_\Psi = \sum\nolimits_{\psi\in\Psi}\sum\nolimits_{k\in[Q]}\mathbf{V}_{\psi,k}\mathbf{P}_{\psi,k}\mathbf{s}_{\psi,k} = \sum\nolimits_{\psi\in\Psi}\mathbf{V}_{\psi}\mathbf{P}_{\psi}\mathbf{s}_{\psi}, \tag{1}$$

where $\mathbf{s}_\psi \in \mathbb{C}^{QM}$ and $\mathbf{V}_\psi \in \mathbb{C}^{L\times QM}$ denote the intended data symbols by the served Q users in the cache-state ψ and the corresponding precoding scheme respectively, and where the diagonal matrix $\mathbf{P}_\psi \in \mathbb{C}^{QM\times QM}$, responsible for power allocation, is obtained by orderly collecting the diagonal elements of the matrices $\{\mathbf{P}_{\psi,k} : \psi \in \Psi, k \in [Q]\}$.

Given $\mathbf{x}_\Psi$ in (1), the received signal vector $\mathbf{y}_{\psi,k} \in \mathbb{C}^M$ at the typical user $\mathrm{U}_{\psi,k}$ takes the form in (2), as shown at the top of next page, where $\mathbf{z}_{\psi,k} \sim \mathcal{CN}(\mathbf{0}_L, N_0\mathbf{I}_L)$ denotes the additive white Gaussian noise (AWGN), and where $\mathbf{H}_{\psi,k} \in \mathbb{C}^{M\times L}$ denotes the channel matrix — from the BS to $\mathrm{U}_{\psi,k}$ — whose elements are independently and identically distributed complex Gaussian random variables with zero-mean and variance $\beta_{\psi,k}$. We note that the factor $\beta_{\psi,k}$ accounts for the large-scale fading and/or pathloss at $\mathrm{U}_{\psi,k}$. As each cache-aided user $\mathrm{U}_{\psi,k}$ knows — as we detail in Section II-A — the messages $\{\mathbf{s}_{\phi,\vartheta} : \phi \in \Psi \setminus \psi, \vartheta \in [Q]\}$ intended by the active users of other user groups in Ψ, we can conclude that the inter-group interference in (2) can be removed by using the cached content in $\mathrm{U}_{\psi,k}$ and the composite CSI $\{\mathbf{H}_{\psi,k}\mathbf{V}_{\phi,\vartheta}\mathbf{P}_{\phi,\vartheta} : \phi \in \Psi \setminus \psi, \vartheta \in [Q]\}$, the cost of which we will account for in our analysis. Specifically, we consider the widely adopted TDD training for the CSI acquisition at both the BS and the users [22], where Θ pilot transmissions per receive antenna are used for CSI acquisition during each channel block corresponding to some coherence time T_c and coherence

[1] We note that this so-called vector coded caching approach in [3], simply 'collapses' (by linearly combining) a carefully selected set of $G = |\Psi|$ vectors into the single vector in (1). Had there been no coded caching, this would have entailed G transmissions one after the other. Therefore, it is conceivable to expect the multiplicative boost to persist for a broader class of precoders.

$$\mathbf{y}_{\psi,k} = \mathbf{H}_{\psi,k}\mathbf{x}_{\Psi} + \mathbf{z}_{\psi,k} = \mathbf{H}_{\psi,k}\mathbf{V}_{\psi,k}\mathbf{P}_{\psi,k}\mathbf{s}_{\psi,k} + \mathbf{z}_{\psi,k} + \underbrace{\mathbf{H}_{\psi,k}\sum_{k'\in[Q]\setminus k}\mathbf{V}_{\psi,k'}\mathbf{P}_{\psi,k'}\mathbf{s}_{\psi,k'}}_{\text{intra-group interference}} + \underbrace{\mathbf{H}_{\psi,k}\sum_{\phi\in\Psi\setminus\psi}\sum_{\vartheta\in[Q]}\mathbf{V}_{\phi,\vartheta}\mathbf{P}_{\phi,\vartheta}\mathbf{s}_{\phi,\vartheta}}_{\text{inter-group interference}}. \quad (2)$$

bandwidth W_c. After removing the inter-group interference via vector coded caching, the signal vector for decoding at $\mathrm{U}_{\psi,k}$ is then of the form

$$\begin{aligned}\mathbf{y}'_{\psi,k} =& \mathbf{H}_{\psi,k}\mathbf{V}_{\psi,k}\mathbf{P}_{\psi,k}\mathbf{s}_{\psi,k}\\ &+ \sum\nolimits_{k'\in[Q]\setminus k}\mathbf{H}_{\psi,k}\mathbf{V}_{\psi,k'}\mathbf{P}_{\psi,k'}\mathbf{s}_{\psi,k'} + \mathbf{z}_{\psi,k}.\end{aligned} \quad (3)$$

Following the conventional ZF precoding for single-antenna receivers in [23], we completely separate the transmitted QM symbol streams to the user-group ψ such that there is no intra-group interference. Therefore, the M symbols simultaneously sent to $\mathrm{U}_{\psi,k}$ are fully separated (corresponding to complete channel diagonalization at the BS), and thus $\mathrm{U}_{\psi,k}$ independently decodes the intended M symbols without interference from other symbols[2]. Specifically, the precoding matrix $\mathbf{V}_{\psi} \in \mathbb{C}^{L\times QM}$ in (1) for the user-group ψ takes the form

$$\mathbf{V}_{\psi} = \left(\mathbf{H}_{\psi}^{H}\left(\mathbf{H}_{\psi}\mathbf{H}_{\psi}^{H}\right)^{-1}\right)\circ\mathbf{N}_{\psi}, \quad (4)$$

where $\circ$ denotes the Hadamard product operator, where $\mathbf{H}_{\psi} \triangleq [\mathbf{H}_{\psi,1}^{T},\cdots,\mathbf{H}_{\psi,Q}^{T}]^{T} \in \mathbb{C}^{QM\times L}$ denotes the channel matrix from the BS to the Q active users of user-group ψ, and where $\mathbf{N}_{\psi} \in \mathbb{C}^{L\times QM}$ is responsible for normalizing the norm-2 of each column in $\mathbf{V}_{\psi}$ into 1. Specifically, the ℓ-th column in $\mathbf{N}_{\psi}$ is $\mathbf{1}_{L}\sqrt{[(\mathbf{H}_{\psi}\mathbf{H}_{\psi}^{H})^{-1}]_{\ell,\ell}}$, where $\mathbf{1}_{L} \in \mathbb{C}^{L}$ is the vector with all elements equaling 1. Given the ZF precoding designed in (4) and under the usual Gaussian signaling assumption, the effective (sum) rate for $\mathrm{U}_{\psi,k}$ takes the form

$$R_{\psi,k}^{\mathrm{ZF}} = \xi_{G,Q}\sum_{m=1}^{M}\ln\left(1+\frac{P_{\psi,k,m}}{N_0\left[\left(\mathbf{H}_{\psi}\mathbf{H}_{\psi}^{H}\right)^{-1}\right]_{k(m),k(m)}}\right), \quad (5)$$

where $k(m) \triangleq (k-1)M+m$, and where $\xi_{G,Q} = 1 - GQM\Theta/T_c/W_c$ accounts for the CSI cost due to TDD training (cf. [22]). For a maximum allowable transmit power P_{tot}, we cast the MMF optimization problem for power allocation as

$$\mathcal{P}_0 : \begin{cases} \max_{\mathbf{P}_{\Psi}}\min_{\psi\in\Psi}\min_{k\in[Q]}\bar{R}_{\psi,k}^{\mathrm{ZF}} \\ \text{s. t. } P_t = \mathbb{E}\{\mathbf{x}_{\Psi}^{H}\mathbf{x}_{\Psi}\} = \mathrm{Tr}\{\mathbf{P}_{\Psi}^{2}\} \le P_{\text{tot}}, \end{cases} \quad (6)$$

where $\bar{R}_{\psi,k}^{\mathrm{ZF}}$ represents the average of $R_{\psi,k}^{\mathrm{ZF}}$ in (5) over channel states.

A. Signal-Level Vector Coded Caching for Finite SNR

Building on the general vector-clique structure in [3], we are here free to choose the precoding schemes, as well as calibrate at will the dimensionality of each vector clique. This freedom is essential in controlling the impact of CSI costs and of power-splitting across users, both of which directly affect the performance in practical SNR regimes [12], [13].

We proceed to describe the cache placement phase and the subsequent delivery phase.

1) Placement Phase: The cache placement policy is from [3]. The first step involves the partition of each library file W_n into $\binom{\Lambda}{\Lambda\gamma}$ non-overlapping equally-sized subfiles $\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\}$, each labeled by some $\Lambda\gamma$-tuple $\mathcal{T} \subseteq [\Lambda]$. As discussed in the introduction, the number of cache states Λ is chosen to satisfy the file-size constraint. Subsequently the K users are *arbitrarily* separated into Λ disjoint groups $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{\Lambda}$, where the g-th cache-group, which consists of $B = \frac{K}{\Lambda}$ users, is given by $\mathcal{D}_g \triangleq \{b\Lambda+g\}_{b=0}^{B-1} \subseteq [K]$. All the users belonging to the same group are assigned the same cache state and thus proceed to cache *identical* content. In particular, for those in the g-th group, this content takes the form $\mathcal{Z}_{\mathcal{D}_g} = \{W_n^{\mathcal{T}} : \mathcal{T} \ni g, \forall n \in [N]\}$.

2) Delivery Phase: This phase starts when each user $\kappa \in [K]$ simultaneously asks for its intended file, denoted here by $W_{d_\kappa}, d_\kappa \in [N]$. The BS selects Q users from each group, where $Q \le B$ is a variable that would have been the multiplexing gain in the cacheless case. By doing so, the BS decides to first 'encode' over the first ΛQ users, and to repeat the encoding process B/Q times. To deliver to the ΛQ users, the transmitter employs $\binom{\Lambda}{\Lambda\gamma+1}$ sequential transmission stages. During each such stage, the BS simultaneously serves a unique set Ψ of $|\Psi| = \Lambda\gamma+1$ groups, corresponding to a total of $Q(\Lambda\gamma+1)$ users served at a time (i.e., per stage), by sending the subfiles $\{W_{d_{\psi,k}}^{\Psi\setminus\psi} : \psi\in\Psi, k\in[Q]\}$ simultaneously. After mapping each such subfile $W_{d_{\psi,k}}^{\Psi\setminus\psi}$ into a complex-valued data vector $\mathbf{s}_{\psi,k}$, the transmission is as shown in (1). The reason of selecting these subfiles is that $\mathrm{U}_{\psi,k}$ has cached the subfiles $\{W_n^{\Psi\setminus\phi} : \phi \in \Psi\setminus\psi, \forall n\in[N]\}$, thereby enabling $\mathrm{U}_{\psi,k}$ to cancel the inter-group interference resulting from $\{W_{d_{\phi,\vartheta}}^{\Psi\setminus\phi} : \phi\in\Psi\setminus\psi, \vartheta\in[Q]\}$. Note that the inter-group interference cancellation also requires knowledge of the composite CSI (cf. (2)). The intra-group interference resulted from $\{W_{d_{\psi,k'}}^{\Psi\setminus\psi} : \psi\in\Psi, k'\in[Q]\setminus k\}$ can be handled with linear precoding that 'separates' the signals of the users from the same group. At the end of the $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, all the ΛQ users obtain their intended files. By repeating this process $\frac{B}{Q}$ times, all the K users obtain their intended files. We refer to [3] for more details.

B. Main Performance Metrics

We will henceforth use the term (G,Q)-vector coded caching, to refer to the vector coded caching scheme when it serves G groups with Q users per group. We will also use the term *ZF-based* (G,Q)*-vector coded caching* to refer to the

[2] We note that the performance difference between this ZF variant and the block-diagonalization precoding (involving receiver output processing) proposed in [24] is marginal in the presence of a small antenna array scale at the users, e.g., $M \le 4$ (cf. [25]).

same scheme when the underlying precoder is the ZF precoder. Let us now formally define some important metrics of interest.

Definition 1. (Effective sum-rate). *For a (G, Q)-vector coded caching scheme, its effective (instantaneous) sum-rate is denoted by $R(G, Q)$ and is defined as the total effective rate (after accounting for CSI costs) summed over the GQ simultaneously served users. Moreover, $\bar{R}(G, Q)$ represents $R(G, Q)$ averaged over channel fading.*

Definition 2. (Effective coded caching gain). *For a given set of SNR, M and L resources, and a fixed underlying precoder class, the effective gain, after accounting for CSI costs, of the (G, Q)-vector coded caching scheme over the cacheless scenario (corresponding to $G = 1$, and an operating multiplexing gain Q'), will be denoted as $\mathcal{G} \triangleq \frac{\bar{R}^\star(G,Q)}{\bar{R}^\star(1,Q')}$, where $\bar{R}^\star(G, Q)$ describes the rate $\bar{R}(G, Q)$ that is optimized via power allocation for MMF (cf. (6)). We also call $\mathcal{G}^\star \triangleq \frac{\max_Q \bar{R}^\star(G,Q)}{\max_{Q'} \bar{R}^\star(1,Q')}$ as the effective gain of optimized rates, where Q and Q' are also independently optimized.*

III. Effective Sum-Rate and Gain Analysis

In this section, we simply perform analysis on the effective sum-rate and the effective gain of ZF-based vector coded caching. More general results are presented in [25].

In the following, we derive the lower and upper bounds on the average of the effective rate in (5), averaged over channel fading.

Proposition 1. *The average effective rate for a typical user $\mathrm{U}_{\psi,k}$ in ZF-based vector coded caching is bounded as $\widetilde{R}^{\mathrm{ZF}}_{\psi,k} \leq \bar{R}^{\mathrm{ZF}}_{\psi,k} \leq \widehat{R}^{\mathrm{ZF}}_{\psi,k}$, where $\widetilde{R}^{\mathrm{ZF}}_{\psi,k}$ and $\widehat{R}^{\mathrm{ZF}}_{\psi,k}$ are respectively given by*

$$\widetilde{R}^{\mathrm{ZF}}_{\psi,k} \triangleq \xi_{G,Q} \sum_{m=1}^{M} \ln \left(1 + \frac{P_{\psi,k,m}(L - QM)\beta_{\psi,k}}{N_0} \right), \tag{7}$$

$$\widehat{R}^{\mathrm{ZF}}_{\psi,k} \triangleq \xi_{G,Q} \sum_{m=1}^{M} \ln \left(1 + \frac{P_{\psi,k,m}(L - QM + 1)\beta_{\psi,k}}{N_0} \right). \tag{8}$$

Proof. The average of the effective rate in (5) over channel states can be written as

$$\begin{aligned}\bar{R}^{\mathrm{ZF}}_{\psi,k} &= \xi_{G,Q}\mathbb{E}\left\{\sum_{m=1}^{M} \ln\left(1 + \frac{P_{\psi,k,m}}{N_0\big[\big(\mathbf{H}_\psi \mathbf{H}_\psi^H\big)^{-1}\big]_{k(m),k(m)}}\right)\right\} \\ &\overset{(a)}{\geq} \xi_{G,Q} \sum_{m=1}^{M} \ln\left(1 + \frac{P_{\psi,k,m}}{N_0 \mathbb{E}\left\{\big[\big(\mathbf{H}_\psi \mathbf{H}_\psi^H\big)^{-1}\big]_{k(m),k(m)}\right\}}\right), \end{aligned} \tag{9}$$

where (a) follows from Jensen's inequality on the convex function $\ln(1 + x^{-1})$. Considering that $\mathbb{E}\big\{\big[\big(\mathbf{H}_\psi^T \mathbf{H}_\psi^*\big)^{-1}\big]_{k(m),k(m)}\big\} = \frac{1}{\beta_{\psi,k}(L-QM)}$ (cf. [26]), we obtain the lower-bound in (7).

To obtain the upper-bound of $\bar{R}^{\mathrm{ZF}}_{\psi,k}$, we use Jensen's inequality for the convex function $\ln(1 + x)$ in the first line of (9), and then we have that

$$\bar{R}^{\mathrm{ZF}}_{\psi,k} \leq \xi_{G,Q} \sum_{m=1}^{M} \ln\left(1 + \frac{P_{\psi,k,m}}{N_0}\mathbb{E}\left\{\frac{1}{\big[\big(\mathbf{H}_\psi \mathbf{H}_\psi^H\big)^{-1}\big]_{k(m),k(m)}}\right\}\right).$$

Using $\mathbb{E}\big\{\big(\big[(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\big]_{k(q),k(q)}\big)^{-1}\big\} = \beta_{\psi,k}(L - QM + 1)$ (cf. [27]) in the above, we can derive (8) □

As performing numerical optimization for MMF in (6) can be computationally prohibitive, especially when having large antenna arrays, we alternatively optimize the bounds in Proposition 1, which leads to the following optimization problems

$$\mathcal{P}_1 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi\in\Psi} \min_{k\in[Q]} \widetilde{R}^{\mathrm{ZF}}_{\psi,k} \\ \text{s. t. } P_t = \mathrm{Tr}\{\mathbf{P}_\Psi^2\} \leq P_{\mathrm{tot}}. \end{cases} \tag{10}$$

$$\mathcal{P}_2 : \begin{cases} \max_{\mathbf{P}_\Psi} \min_{\psi\in\Psi} \min_{k\in[Q]} \widehat{R}^{\mathrm{ZF}}_{\psi,k} \\ \text{s. t. } P_t = \mathrm{Tr}\{\mathbf{P}_\Psi^2\} \leq P_{\mathrm{tot}}. \end{cases} \tag{11}$$

Obviously, P_t should reach its upper-bound P_{tot} when the optimum in (10) and (11) is achieved. We note that the power allocation policy on $\{s_{\psi,k,m} : m \in [M]\}$ does not affect the effective rates for other served users as there is no inter-user interference after ZF precoding. In accordance with the water-filling algorithm (cf. [28, Ch. 10]), it is easy to see that equal power allocation among $\{s_{\psi,k,m} : m \in [M]\}$ for $\forall\psi \in \Psi$ and $k \in [Q]$ is optimal for both $\widetilde{R}^{\mathrm{ZF}}_{\psi,k}$ and $\widehat{R}^{\mathrm{ZF}}_{\psi,k}$, i.e., $P_{\psi,k,m} = P_{\psi,k}/M$, where $P_{\psi,k} = \mathrm{Tr}\{\mathbf{P}_{\psi,k}\}$ denotes the total power allocated to $\mathrm{U}_{\psi,k}$. To solve the MMF problems in (10) and (11), we have the following theorem.

Theorem 1. *The optimal MMF-constrained (average) effective sum-rate $\bar{R}^\star_{\mathrm{ZF}}$ is bounded as $\widetilde{R}^\star_{\mathrm{ZF}} \leq \bar{R}^\star_{\mathrm{ZF}} \leq \widehat{R}^\star_{\mathrm{ZF}}$, where $\widetilde{R}^\star_{\mathrm{ZF}}$ and $\widehat{R}^\star_{\mathrm{ZF}}$ are respectively*[3]

$$\widetilde{R}^\star_{\mathrm{ZF}} = \xi_{G,Q} GQM \ln\left(1 + \frac{P_{\mathrm{tot}}(L - QM)}{MN_0 \sum_{\psi\in\Psi}\sum_{k\in[Q]} \beta^{-1}_{\psi,k}}\right),$$

$$\widehat{R}^\star_{\mathrm{ZF}} = \xi_{G,Q} GQM \ln\left(1 + \frac{P_{\mathrm{tot}}(L - QM + 1)}{MN_0 \sum_{\psi\in\Psi}\sum_{k\in[Q]} \beta^{-1}_{\psi,k}}\right).$$

Proof. We first note that when the optimum for $\mathcal{P}_1$ is achieved, all the simultaneously served users will have the same[4] lower-bound for their average effective rates (i.e., $\widetilde{R}^\star_{\mathrm{ZF}}/G/Q$). Based on the lower-bound expression in (7) and water-filling algorithm, we can express the power allocated to $\mathrm{U}_{\psi,k}$ to achieve $\widetilde{R}^\star_{\mathrm{ZF}}/G/Q$ as

$$P^\star_{\psi,k} = \left(\exp\left(\frac{\widetilde{R}^\star_{\mathrm{ZF}}}{\xi_{G,Q}GQM}\right) - 1\right)\frac{N_0}{(L - QM)\beta_{\psi,k}}. \tag{12}$$

By using the total power constraint $\sum_{\psi\in\Psi}\sum_{k\in[Q]} P_{\psi,k} = P_{\mathrm{tot}}$, we finally derive the expression for $\widetilde{R}^\star_{\mathrm{ZF}}$. The derivation of $\widehat{R}^\star_{\mathrm{ZF}}$ follows in a similar manner. □

[3] It is obvious that the gap between $\widetilde{R}^\star_{\mathrm{ZF}}$ and $\widehat{R}^\star_{\mathrm{ZF}}$ will be negligible for $L \gg QM$. Moreover, both the lower and upper bounds in Theorem 1 depend on the cumulative effect of the inverse pathloss in all served users.

[4] To see this, let us consider its opposite, and let us note that when we arrive at the optimal power allocation for $\mathcal{P}_1$, if there exists one user whose effective rate is higher than the smallest rate, this user can 'borrow' some power to the user with the smallest rate until their rates are the same, without affecting the rates for other users, which enhances the smallest effective rates, and which is contradictory to the optimal power allocation assumption.

With the help of Theorem 1, we are enabled to bound the effective gain of optimized rates as $\widetilde{\mathcal{G}}^{\star}_{\mathrm{ZF}} \leq \mathcal{G}^{\star}_{\mathrm{ZF}} \leq \widehat{\mathcal{G}}^{\star}_{\mathrm{ZF}}$, where $\widetilde{\mathcal{G}}^{\star}_{\mathrm{ZF}}$ and $\widehat{\mathcal{G}}^{\star}_{\mathrm{ZF}}$ are defined respectively as

$$\widetilde{\mathcal{G}}^{\star}_{\mathrm{ZF}} \triangleq \frac{\max\limits_{Q\in[Q_{\max}]} \widetilde{R}^{\star}_{\mathrm{ZF}}(G,Q)}{\max\limits_{Q'\in[Q_{\max}]} \widehat{R}^{\star}_{\mathrm{ZF}}(1,Q')}, \quad \widehat{\mathcal{G}}^{\star}_{\mathrm{ZF}} \triangleq \frac{\max\limits_{Q\in[Q_{\max}]} \widehat{R}^{\star}_{\mathrm{ZF}}(G,Q)}{\max\limits_{Q'\in[Q_{\max}]} \widetilde{R}^{\star}_{\mathrm{ZF}}(1,Q')},$$

where $Q_{\max} = \frac{L-1}{M}$ is the maximum allowable multiplexing gain in ZF precoding. In the following corollary, we parameterize the lower and upper bounds of the gain $\mathcal{G}^{\star}_{\mathrm{ZF}}$ of vector coded caching.

Corollary 1. *Given ZF-based precoding at the BS and given M-antenna receivers, the lower and upper bounds of the optimal effective gain $\mathcal{G}^{\star}_{\mathrm{ZF}}$ are respectively*

$$\widetilde{\mathcal{G}}^{\star}_{\mathrm{ZF}} = \frac{\max\limits_{Q\in[Q_{\max}]} \xi_{G,Q} GQM \ln\left(1 + \frac{P_{\mathrm{tot}}(L-QM)}{MN_0 \sum_{\psi\in\Psi}\sum_{k\in[Q]} \beta^{-1}_{\psi,k}}\right)}{\max\limits_{Q'\in[Q_{\max}]} \xi_{1,Q'} Q'M \ln\left(1 + \frac{P_{\mathrm{tot}}(L-Q'M+1)}{MN_0 \sum_{k'\in[Q']} \beta^{-1}_{k'}}\right)},$$

$$\widehat{\mathcal{G}}^{\star}_{\mathrm{ZF}} = \frac{\max\limits_{Q\in[Q_{\max}]} \xi_{G,Q} GQM \ln\left(1 + \frac{P_{\mathrm{tot}}(L-QM+1)}{MN_0 \sum_{\psi\in\Psi}\sum_{k\in[Q]} \beta^{-1}_{\psi,k}}\right)}{\max\limits_{Q'\in[Q_{\max}]} \xi_{1,Q'} Q'M \ln\left(1 + \frac{P_{\mathrm{tot}}(L-Q'M)}{MN_0 \sum_{k'\in[Q']} \beta^{-1}_{k'}}\right)}.$$

Proof. It is easy to derive Corollary 1 by using Theorem 1 and Definition 2. □

IV. Numerical Results

This section presents various numerical results that validate our analysis and provide insightful comparisons. We consider relatively low mobility users, and assume that $T_c = 0.05$ s and $W_c = 300$ kHz, corresponding to a coherence block of 15000 symbols. We consider CSI pilot length of $\Theta = 10$, which is expected to be sufficient for providing near-perfect CSI at both the BS and the users [22]. The AWGN spectral density is considered to be -174 dBm/Hz, and the spectrum bandwidth for each user is 20 MHz. We generate 1000 realizations of user locations, based on the assumption of uniformly-distributed users across a single cell. We consider a Macro-cell with an inner radius of 35 meters and an outer radius of 500 meters, as well as consider a Micro-cell with an inner radius of 10 meters and an outer radius of 100 meters. Assuming a carrier frequency of 2 GHz, in the Macro-cell case, the pathloss is modeled as $\beta_{\psi,k} = l_0 r^{-\eta}_{\psi,k}$, where $r_{\psi,k}$ is the distance between the BS and $\mathrm{U}_{\psi,k}$, where $\eta = 3.76$ is the pathloss exponent, and where $l_0 = 10^{-3.53}$ regulates the channel attenuation at 35 meters [29]. For the Micro-cell scenario, we note that the pathloss model can often differ when considering delivery distance between 10 and 40 meters, compared to when considering a distance in the [40, 100] meters range. For simplicity though, we here use the pathloss model for [10, 40] meters over the entire delivery range in the Micro-cell, and thus consider $l_0 = 10^{-3.7}$ and $\eta = 3$ (cf. [29, Table II]).

The sole purpose of Fig. 1 is to exhibit the tightness of using the presented bounds instead of the actual expressions. This figure plots the average effective sum-rate of vector coded caching, where the sum-rate is optimized via power allocation for MMF. In this same figure, the simulated results are created by using the built-in function 'fminimax' in MATLAB for directly solving $\mathcal{P}_0$ in (6), while the lower-bound and upper-bound respectively correspond to $\widetilde{R}^{\star}_{\mathrm{ZF}}$ and $\widehat{R}^{\star}_{\mathrm{ZF}}$ in Theorem 1. It is obvious that both the upper and lower bounds almost match the simulated actual result over the entire P_{tot} range.

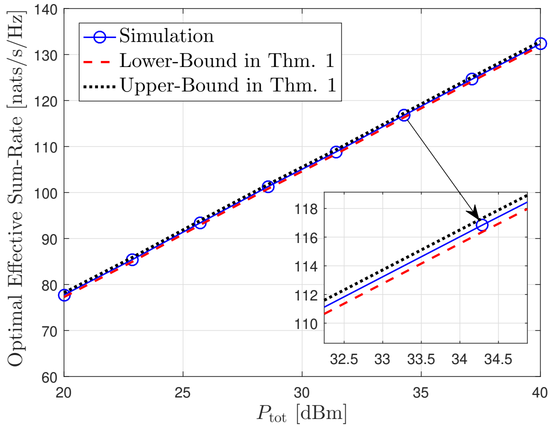

Fig. 1: Bounded Rates of Vector Coded Caching for $L = 16$, $G = 3$, and $Q = M = 2$ in a Micro-cell

In Figs. 2–3, we numerically evaluate the lower and upper bounds in Theorem 1, which have been shown to be tight even for a modest array sizes (see Fig. 1). In these Figs. 2–3, vector coded caching and the cacheless scenario share the same system parameters (e.g., L and M), where though, very importantly, the corresponding multiplexing gains Q and Q' are independently optimized under the assumption of Λ dividing K. We note that the impact of a non-integer $B = K/\Lambda$ on the effective coded caching gain is marginal [13]. For clarity in plotting, we omit numerical results for the cacheless sum-rate in Figs. 2–3.

It is worth noting that Fig. 2 shows notable gains in the average effective sum-rate over the cacheless setting for realistic values of transmit power in a Micro-cell (e.g., $\mathcal{G}^{\star}_{\mathrm{ZF}} \geq 400\%$ when $P_{\mathrm{tot}} \geq 30$ dBm). These gains are somewhat reduced (see Fig. 3) in the Macro-cell (e.g., $\mathcal{G}^{\star}_{\mathrm{ZF}} = 200\%$ for a realistic transmit power of 40 dBm) due to a much larger coverage area which implies a heavy near-far effect, despite allowing for optimal power allocation for MMF. We can also see that increasing the number of transmit antennas from 64 to 128 maintains or even slightly enhances the effective gains in the medium to high P_{tot} region in the Macro-cell. Remedies for further improving the effective gain in Macro-cells will be investigated in our future work.

V. Conclusions

We have investigated ZF-based vector coded caching under various realistic considerations such as having variable pathloss, multi-antenna receivers, and MMF. Specifically, we have derived closed-form expressions for the lower and upper bounds of the effective sum-rate and of the effective gain. Numerical results have validated very clearly the tightness of the derived bounds, revealing again notable effective gains that vector

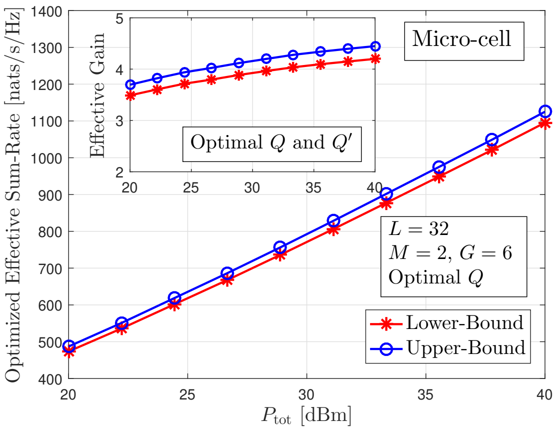

Fig. 2: Optimized Rate and Vector Coded Caching Gain (Micro-cell)

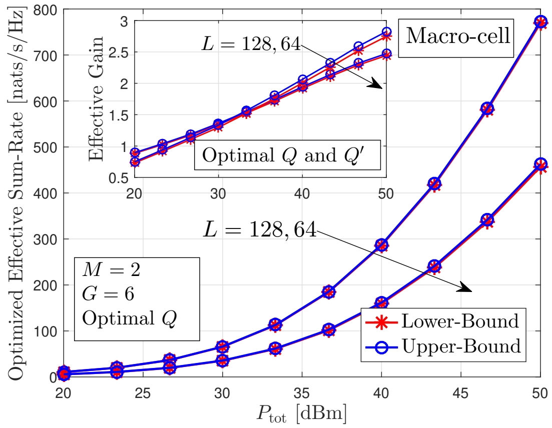

Fig. 3: Optimized Rate and Vector Coded Caching Gain (Macro-cell)

coded caching can provide in realistic scenarios. These gains are particularly high in Micro-cell environments. For example, we have seen that vector coded caching can offer more than a 400% boost in the overall throughput over the traditional (cacheless) MU-MIMO system, under the aforementioned realistic assumptions. We note that all recorded gains attributed to vector coded caching, are taken to reflect the improvement over *optimized* traditional MU-MIMO systems, where for example, such traditional MU-MIMO system is optimized w.r.t. the operational multiplexing gain. We believe that these results serve as additional evidence that vector coded caching can play an important role as an appendix to existing MU-MIMO systems.

References

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

[2] Q. Yan, M. Cheng, X. Tang, and Q. Chen, “On the placement delivery array design for centralized coded caching scheme,” *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5821–5833, Sep. 2017.

[3] E. Lampiris and P. Elia, “Adding transmitters dramatically boosts coded-caching gains for finite file sizes,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[4] S. P. Shariatpanahi *et al.*, “Multi-server coded caching,” *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[5] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Fundamental limits of cache-aided interference management,” *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[6] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, “Physical-layer schemes for wireless coded caching,” *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[7] K.-H. Ngo, S. Yang, and M. Kobayashi, “Scalable content delivery with coded caching in multi-antenna fading channels,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.

[8] M. Bayat, R. K. Mungara, and G. Caire, “Achieving spatial scalability for coded caching via coded multipoint multicasting,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 227–240, Jan. 2019.

[9] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, “Multi-antenna interference management for coded caching,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

[10] H. B. Mahmoodi, M. Salehi, and A. Tölli, “Non-symmetric coded caching for location-dependent content delivery,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),*, Jul. 2021, pp. 712–717.

[11] “The industry’s first independent benchmark study of 5G NR MU-MIMO,” Signals Research Group, Tech. Rep., Sept. 2020.

[12] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Vector coded caching greatly enhances massive MIMO,” in *Proc. 23rd IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.

[13] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Vector coded caching multiplicatively boosts the throughput of realistic downlink systems,” *IEEE Trans. Wireless Commun.*, accepted, doi: 10.1109/TWC.2022.3213475.

[14] H. Zhao, E. Lampiris, G. Caire, and P. Elia, “Multi-antenna coded caching analysis in finite SNR and finite subpacketization,” in *Proc. Int. ITG Workshop on Smart Antennas (WSA)*, Nov. 2021, pp. 433–438.

[15] M. Ji and R.-R. Chen, “Caching and coded multicasting in slow fading environment,” in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[16] B. Tegin and T. M. Duman, “Coded caching with user grouping over wireless channels,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 920–923, Jun. 2020.

[17] H. Zhao, A. Bazco-Nogueras, and P. Elia, “Wireless coded caching with shared caches can overcome the near-far bottleneck,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 350–355.

[18] H. Zhao, A. Bazco-Nogueras and P. Elia, “Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5450–5466, Jul. 2022.

[19] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, “Coded caching and spatial multiplexing gains in MIMO interference networks,” in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[20] Y. Cao and M. Tao, “Treating content delivery in multi-antenna coded caching as general message sets transmission: A DoF region perspective,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3129–3141, Jun. 2019.

[21] A. Destounis, M. Kobayashi, G. Paschos, and A. Ghorbel, “Alpha fair coded caching,” in *Proc. Int. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017, pp. 1–8.

[22] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, “Multiuser MIMO achievable rates with downlink training and channel state feedback,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.

[23] G. Caire and S. Shamai, “On the achievable throughput of a multi-antenna Gaussian broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 43, no. 7, pp. 1691–1706, Jul. 2003.

[24] Q. Spencer, A. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels,” *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[25] H. Zhao, “High performance cache-aided downlink systems: Novel algorithms and analysis,” Ph.D. dissertation, Sorbonne University, 2022.

[26] A. M. Tulino and S. Verdú, “Random matrix theory and wireless communications,” *Foundations Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, Jun. 2004.

[27] K. K. Wong and Z. Pan, “Array gain and diversity order of multiuser MISO antenna systems,” *Int. J. Wireless Inf. Netw.*, vol. 15, no. 2, pp. 82–89, May 2008.

[28] A. Goldsmith, *Wireless Communications.* Cambridge University Press, 2005.

[29] E. Björnson, M. Kountouris, and M. Debbah, “Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination,” in *Proc. Int. Conf. Telecommun. (ICT)*, May 2013, pp. 1–5.