Vector Coded Caching Greatly Enhances Massive MIMO

Hui Zhao*, Antonio Bazco-Nogueras†, and Petros Elia*
*Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France
†IMDEA Networks Institute, 28918 Madrid, Spain
Email: hui.zhao@eurecom.fr; antonio.bazco@imdea.org; elia@eurecom.fr

***Abstract*—The use of vector coded caching has been shown to provide important gains and, more importantly, to alleviate the impact of the file-size constraint, which prevents coded caching from obtaining its ideal gains in practical settings. In this work, we analyze the performance of vector coded caching in the massive MIMO regime, aiming at understanding the benefits that allowing users to cache a practical amount of data could bring to realistic settings in such massive MIMO regime. In particular, we separately consider two linear precoding schemes and analyze the corresponding throughput, for which we derive simple but precise upper and lower bounds. These bounds enable us to characterize the delivery speed-up gain over the uncoded caching setting when the CSI acquisition costs are taken into account. Numerical results demonstrate the tightness of the derived bounds and show a significant boost over uncoded caching and the standard cacheless setting.**

***Index Terms*—Coded caching, linear precoding, massive MIMO.**

I. Introduction

Coded caching was shown to yield a multiplicative delivery speed-up of $K\gamma + 1$ over uncoded caching in a single-stream and error-free shared Broadcast Channel [1], where K and γ represent the total number of served users and the normalized cache size at each user, respectively. This speed-up factor of $K\gamma+1$ is also the degrees-of-freedom (DoF) provided by coded caching in [1]. However, this seminal coded caching framework [1] requires the file size to grow exponentially with K. Under realistic file-size constraints, the DoF of $K\gamma + 1$ is reduced to $\Lambda\gamma + 1$, where Λ is obtained from some maximum allowable subpacketization level S, such that $\Lambda = \arg\max_\lambda \binom{\lambda}{\lambda\gamma}$ s.t. $\binom{\lambda}{\lambda\gamma} \leq S$. In practice, $\Lambda \ll K$ and, more precisely, $\Lambda\gamma$ remains in the single-digit range [2], [3].

As multiple-input multiple-output (MIMO) (and in particular massive MIMO) framework has been a key technology in the current development of wireless networks [4], [5], the implementation of coded caching should be incorporated in such MIMO systems, which leads to the development of *multi-antenna coded caching* [6]–[10]. However, for the first proposed solutions, multi-antenna coded caching was providing a theoretic DoF of $\Lambda\gamma+Q$, where Q is the multiplexing gain provided by the multiple antennas. In the massive MIMO regime, when $\Lambda\gamma \ll Q$, the DoF gain due to caching becomes marginal. Fortunately, this situation was reversed after introducing the so-called *vector coded caching* [11]. Specifically, vector coded caching [11] provides a DoF of $Q(\Lambda\gamma+1)$ while also achieving a dramatically reduced subpacketization, since the file size grows exponentially in Λ/Q.

This work is supported in part by the European Research Council under the EU Horizon 2020 research and innovation program/ERC grant agreement no. 725929 (ERC project DUALITY), and by the Regional Government of Madrid through the grant 2020-T2/TIC-20710 for Talent Attraction.

However, [11] focused on the importance of vector coded caching in the high-SNR (DoF) sense and did not consider any practical issue, such as realistic SNR ranges, impact of beamforming, or performance costs from gathering channel state information (CSI). With the exception of some numerical beamforming optimization in the preliminary work [12], the real performance of such approach compared to uncoded caching still remains unknown.

The multiplicative theoretic DoF boost $\Lambda\gamma + 1$ that vector coded caching provides over its uncoded caching counterpart with multiplexing gain Q can be helpful to effectively alleviate the network congestion generated from the ever-increasing user density [13]. This motivates us to investigate the real performance of vector coded caching in a practical system. In this paper, we analyze the performance of vector coded caching [11] in the massive MIMO regime by considering realistic SNR values, practical linear precoders, effects of beamforming, and CSI costs. The main contributions are outlined as follows.

- We derive simple closed-form upper and lower bounds of the effective sum-rate (later defined in Definition 1) of vector coded caching under two practical linear precoding schemes: Matched Filter (MF) and Zero-Forcing (ZF).
- Making use of the derived upper and lower bounds, we obtain a lower-bound of the effective coded caching gain. Numerical results demonstrate the tightness of this lower-bound and show a significant effective gain over standard linear transmission without coded caching.

Notation: $\mathbb{C}$ stands for set of complex numbers, $\mathbf{I}_L \in \mathbb{C}^{L\times L}$ denotes the identity matrix, and $\mathbf{0}_L \in \mathbb{C}^L$ denotes the all-zeros vector. $|\cdot|$ denotes the cardinality of a set or the absolute value of a complex number, $||\cdot||$ denotes the norm-2 operator for a vector, and we define $[Z] \triangleq \{1, ..., Z\}$ for a positive integer Z. $\mathrm{Tr}\{\cdot\}$ and $\mathbb{E}\{\cdot\}$ denote the trace and the expectation operators, whereas $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$ are the non-conjugate transpose, conjugate, and conjugate transpose of a matrix, respectively.

II. System Model and Problem Description

We consider a downlink scenario where an L-antenna base station (BS) serves K single-antenna users. The BS has access to a library of N equally-sized files, and each user is endowed with a cache that can store a fraction $\gamma \in [0, 1]$ of the library content $\{W_n\}_{n=1}^N$, where W_n denotes the n-th library file.

We consider a symmetric Rayleigh fading channel, where all channel coefficients are assumed to be independent and identically distributed (i.i.d.). Upon denoting the set of users

to which the transmitted signal is intended as $\mathcal{K} \subseteq [K]$, the received signal at any user $k \in \mathcal{K}$ is given by $y_k = \mathbf{h}_k^T \mathbf{x}_{\mathcal{K}} + z_k$, where $z_k \in \mathbb{C}$ represents the corresponding Additive White Gaussian Noise (AWGN) with zero-mean and unit-variance, $\mathbf{x}_{\mathcal{K}} \in \mathbb{C}^L$ denotes the transmitted signal vector that simultaneously serves the users in $\mathcal{K}$, and where $\mathbf{h}_k \in \mathbb{C}^L$ represents the channel vector between the BS and user k. As mentioned, $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$. Finally, $\mathbf{x}_{\mathcal{K}}$ is obtained by applying a specific precoding scheme (which we will detail later on) to the information vector $\mathbf{s}_{\mathcal{K}} \in \mathbb{C}^{|\mathcal{K}|}$ intended for the users in $\mathcal{K}$, where $\mathbf{s}_{\mathcal{K}}$ has mean $\mathbf{0}_{|\mathcal{K}|}$ and covariance matrix $\mathbf{I}_{|\mathcal{K}|}$.

A. Signal-Level Vector Coded Caching for Finite SNR

The coded caching transmission here considered builds on the general vector-clique structure presented in [11], but we are here allowed to select different precoding schemes and variate the dimensionality of each vector clique. This added flexibility is crucial in optimizing the performance when both CSI costs and power-splitting across users are considered, since both aspects impact the performance in practical SNR regimes [14]. We proceed to describe the cache placement phase and the subsequent delivery phase.

1) Placement Phase: The system assumes Λ different *cache states* (i.e., different non-disjoint subsets of the library content). The number of cache states Λ is chosen to satisfy the file-size constraint. The first step involves the partition of each library file W_n into $\binom{\Lambda}{\Lambda\gamma}$ non-overlapping equally-sized subfiles $\left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\right\}$, each labeled by some $\Lambda\gamma$-tuple $\mathcal{T} \subseteq [\Lambda]$. Subsequently the K users are *uniformly* distributed into Λ disjoint groups $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\Lambda$, where the g-th group consists of $B \triangleq \frac{K}{\Lambda}$ users and is given by $\mathcal{D}_g \triangleq \left\{b\Lambda + g\right\}_{b=0}^{B-1} \subseteq [K]$. The ϑ-th user of this g-th group is denoted[1] by $\mathrm{U}_{g,\vartheta}$.

At this point, 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 $\mathcal{Z}_{\mathcal{D}_g} = \left\{W_n^{\mathcal{T}} : \mathcal{T} \ni g, \forall n \in [N]\right\}$.

2) Delivery Phase: This phase starts when each user $\kappa \in [K]$ simultaneously asks for its intended file, denoted here by W_{d_κ}, $d_\kappa \in [N]$. The BS selects Q users from each group, where $Q \in [B]$ is the equivalent of the multiplexing gain. By doing so, the BS decides to first 'encode' over the first ΛQ users, and to repeat the encoding process B/Q times.[2]

To deliver to the first 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). At the end of the $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, all the ΛQ users obtain their intended files. As suggested above, the factor $G \triangleq \Lambda\gamma + 1$ describes the number of user-groups that are simultaneously served.

Let us focus on a single transmission stage, where a set $\Psi \subseteq \Lambda$ of $G = \Lambda\gamma + 1$ groups is selected and we serve $Q \leq B$ users per group. In particular, for each user $\mathrm{U}_{\psi,\vartheta}$ of some group $\psi \in \Psi$, this stage delivers all subfiles[3] $s_{\psi,\vartheta}$ by transmitting

$$\mathbf{x}_\Psi = \sum\nolimits_{\psi \in \Psi} \tfrac{\rho_\psi}{\sqrt{G}} \mathbf{V}_\psi \mathbf{s}_\psi, \tag{1}$$

where $\mathbf{V}_\psi \triangleq [\mathbf{v}_{\psi,1} | \ldots | \mathbf{v}_{\psi,Q}]$ and $\mathbf{v}_{\psi,\vartheta} \in \mathbb{C}^{L \times 1}$ denotes the precoder applied to the subfile intended by user $\mathrm{U}_{\psi,\vartheta}$, whereas $\mathbf{s}_\psi \triangleq [s_{\psi,1}, \ldots, s_{\psi,Q}]^T$, and where ρ_ψ denotes the power normalization factor for group $\psi \in \Psi$, applied under a total power constraint P_t.

This scheme just simultaneously delivers a carefully selected linear combination of G linear-precoding vectors. Moreover, the above scheme also incorporates the traditional cacheless downlink scenario (i..e, $\gamma = 0$) which corresponds to $G = 1$. In such case, we obtain the simpler usual expression $\mathbf{x} = \rho \mathbf{V} \mathbf{s}$.

For decoding to work, the subfiles must be chosen carefully following the principles of vector coded caching, such that for the transmission stage which serves the group-set Ψ, the subfile transmitted to user $\mathrm{U}_{\psi,\vartheta}$ is here selected to be $W_{d_{\psi,\vartheta}}^{\Psi \setminus \{\psi\}}$, because this subfile is stored in the cache of each user of every other group in Ψ except ψ. Due to this structure, the users can remove the *inter-group* interference from the other $\Lambda\gamma$ groups by using their cached content. On the other hand, the *intra-group* interference is handled with linear precoding.

Certainly, both the interference nulling and the cache-aided removal of interference require precise estimates of the composite precoder-channel coefficients (cf. (3), (4)), and we will account for these so-called *composite CSI* costs.

B. Vector Coded Caching for the Physical Layer

As is common in practical massive MIMO, we assume TDD uplink-downlink transmission, such that the BS and the users estimate the downlink channels through pilot transmissions by applying channel reciprocity [4].

Let us focus now on describing the transmission that serves a specific set Ψ of user-groups. We consider that there exists a nominal power constraint, which is denoted by P_t. Then, the power normalization factor ρ_ψ from (1) takes the form

$$\rho_\psi = P_t^{1/2} \left(\mathrm{Tr}\{\mathbf{V}_\psi^H \mathbf{V}_\psi\}\right)^{-1/2}. \tag{2}$$

The subsequent corresponding received signal at user $\mathrm{U}_{\psi,k}$ (i.e., at the k-th user of group $\psi \in \Psi$), will take the form

$$y_{\psi,k} = \tfrac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \rho_\psi \mathbf{V}_\psi \mathbf{s}_\psi + \underbrace{\tfrac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \sum_{\phi \in \Psi, \phi \neq \psi} \rho_\phi \mathbf{V}_\phi \mathbf{s}_\phi}_{\text{inter-group interference}} + z_{\psi,k}. \tag{3}$$

As previously mentioned, this inter-group interference can be removed from $y_{\psi,k}$ by exploiting that same user's cached content and that user's composite CSI $\{\mathbf{h}_{\psi,k}^T \mathbf{v}_{\phi,k'} \rho_\phi\}_{\phi \in \{\Psi \setminus \psi\},\, k' \in [Q]}$.

[1] We will henceforth consider K to be a multiple of Λ for the sake of clarity of exposition, which does not limit the scope of the results in any way. The general case can be readily handled (cf. [11]). Furthermore, this grouping and the entire placement phase are naturally done before users' requests are made and before the channel states are known to the BS.

[2] This means that, if there exist $B = aQ$ users per group ($K = a\Lambda Q$), then the algorithm will be applied to the first ΛQ users, and, once the delivery to these users is finished, the same algorithm will be independently applied to the next ΛQ users, repeating the process a times until all K users are served.

[3] In a slight abuse of notation, we use the term "subfile" to refer both to the actual subfile generated after file-splitting, as well as to the corresponding complex-valued information symbol $s_{\psi,\vartheta}$.

Then, after the cache-aided removal of this inter-group interference, the equivalent received signal at $\mathrm{U}_{\psi,k}$ is given by

$$y'_{\psi,k}=\frac{\rho_\psi}{\sqrt{G}}(\mathbf{h}_{\psi,k}^T\mathbf{v}_{\psi,k}s_{\psi,k}+\underbrace{\sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\mathbf{h}_{\psi,k}^T\mathbf{v}_{\psi,\vartheta}s_{\psi,\vartheta}}_{\text{intra-group interference}})+z_{\psi,k}. \quad (4)$$

Consequently, under the usual Gaussian signaling assumption, the corresponding SINR for information decoding at $\mathrm{U}_{\psi,k}$ is

$$\mathrm{SINR}_{\psi,k}=\frac{\rho_\psi^2|\mathbf{h}_{\psi,k}^T\mathbf{v}_{\psi,k}|^2}{G}\big(1+\frac{\rho_\psi^2}{G}\sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}|\mathbf{h}_{\psi,k}^T\mathbf{v}_{\psi,\vartheta}|^2\big)^{-1}. \quad (5)$$

Next, we present the *effective transmission rate*, i.e., the actual data rate after accounting for the need for CSI sharing [15], [16]. For user $\mathrm{U}_{\psi,k}$, it reads as

$$R_{\psi,k}\triangleq\xi_{GQ}\ln\left(1+\mathrm{SINR}_{\psi,k}\right), \quad (6)$$

where $\xi_{GQ}\triangleq(1-\frac{\Theta GQ}{T})$ accounts for the CSI costs due to TDD training, since only $T-\Theta GQ$ symbols are used for user data transmission; T is the coherence block length (in symbols), and Θ is the number of resources per user and per block used for pilot transmission (cf. [15], [16]).

We consider the MF and ZF linear precoding schemes, which are selected for its simplicity, usage and competitiveness in terms of performance [4]. Hence, we have the following cases:

$$\mathbf{V}_\psi=\begin{cases}\mathbf{H}_\psi^H & \text{MF Precoder}\\ \mathbf{H}_\psi^H\left(\mathbf{H}_\psi\mathbf{H}_\psi^H\right)^{-1} & \text{ZF Precoder,}\end{cases} \quad (7)$$

where $\mathbf{H}_\psi\triangleq\left[\mathbf{h}_{\psi,1}|\mathbf{h}_{\psi,2}|\cdots|\mathbf{h}_{\psi,Q}\right]^T\in\mathbb{C}^{Q\times L}$ denotes the channel matrix between the BS and the Q chosen users.

We 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 also use the term *MF-based (resp. ZF-based)* (G,Q)*-vector coded caching* to refer to the same scheme when the underlying precoder is MF (resp. ZF). Let us formally define two important metrics of interest.

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

Definition 2. (Effective coded caching gain). *The effective gain after accounting for CSI costs of the (G,Q)-vector coded caching over the cacheless/uncoded caching*[4] *scenario (corresponding to $G=1$, cf. [11]) with operating multiplexing gain Q' is defined as $\mathcal{G}(G,Q;1,Q')\triangleq\frac{\bar{R}(G,Q)}{\bar{R}(1,Q')}$.*

III. Large-Scale Antenna Analysis

In this section, we analyze the effective sum-rate and the effective coded caching gain of the scheme presented in Section II-A for MF and ZF precoders. Specifically, we analyze the regime where the number of transmit antennas L grows unboundedly while Q remains constant, i.e., where the number of simultaneously served users remains fixed for a given value of K,γ. The analysis of the other asymptotic regime, where $L,Q\to\infty$ while $\frac{Q}{L}$ remains fixed, is presented in [17].

A. MF Precoding Analysis

Before presenting the main results on the rate and effective gain for MF precoding, we focus on the power factor ρ_ψ.

From (2) and (7), we have that $\rho_\psi^2=\frac{P_t}{\mathrm{Tr}\{\mathbf{H}_\psi\mathbf{H}_\psi^H\}}$. Note that $\mathrm{Tr}\{\frac{1}{L}\mathbf{H}_\psi\mathbf{H}_\psi^H\}=\frac{1}{L}\sum_{k=1}^Q\mathbf{h}_{\psi,k}^T\mathbf{h}_{\psi,k}^*$ converges to Q almost surely $(a.s.)$ as $L\to\infty$ according to the Strong Law of Large Numbers. Thus, we have that $\rho_\psi\xrightarrow{a.s.}\sqrt{P_t/(LQ)}$.

Since we are interested in the large number of antennas regime, we consider for the sake of clarity that $\rho_\psi=\sqrt{\frac{P_t}{LQ}}$ for MF precoding. Then, the SINR at $\mathrm{U}_{\psi,k}$ becomes

$$\mathrm{SINR}_{\psi,k}^{\mathrm{MF}}=\frac{\frac{P_t}{GQL}\left|\mathbf{h}_{\psi,k}^T\mathbf{h}_{\psi,k}^*\right|^2}{1+\frac{P_t}{GQL}\sum_{\vartheta=1,\vartheta\neq k}^Q\left|\mathbf{h}_{\psi,k}^T\mathbf{h}_{\psi,\vartheta}^*\right|^2}. \quad (8)$$

Let us present our first main result, which focuses on the average rate of the proposed scheme with MF precoding.

Lemma 1. *For any fixed $Q>2$ and $L\to\infty$, the effective sum-rate of MF-based vector coded caching, denoted by $\bar{R}_{\mathrm{sum}}^{\mathrm{MF}}$, is bounded by $\widetilde{R}_{\mathrm{sum}}^{\mathrm{MF}}\leq\bar{R}_{\mathrm{sum}}^{\mathrm{MF}}\leq\widehat{R}_{\mathrm{sum}}^{\mathrm{MF}}$, where*

$$\widetilde{R}_{\mathrm{sum}}^{\mathrm{MF}}\triangleq\xi_{GQ}GQ\ln\left(1+\frac{P_t(L-1)(L-2)}{GQL+P_t(Q-1)(L-2)}\right) \quad (9)$$

$$\widehat{R}_{\mathrm{sum}}^{\mathrm{MF}}\triangleq\xi_{GQ}GQ\ln\left(\frac{GQ+P_t(L+Q)}{GQ+P_t(Q-2)(L-1)/L}\right) \quad (10)$$

Proof. The proof is relegated to Appendix I. □

Corollary 1. *For any fixed $Q,Q'>2$ and $L\to\infty$, the effective gain of MF-based (G,Q)-vector coded caching, denoted by $\mathcal{G}_{\mathrm{MF}}$, can be lower-bounded by*

$$\mathcal{G}_{\mathrm{MF}}\geq\frac{\widetilde{R}_{\mathrm{sum}}^{\mathrm{MF}}(G,Q)}{\widehat{R}_{\mathrm{sum}}^{\mathrm{MF}}(1,Q')}=\frac{\xi_{GQ}GQ}{\xi_{Q'}Q'}\frac{\ln\left(1+\frac{P_t(L-1)(L-2)}{GQL+P_t(Q-1)(L-2)}\right)}{\ln\left(\frac{GQ'+P_t(L+Q')}{GQ'+P_t(Q'-2)(L-1)/L}\right)}.$$

Proof. Corollary 1 can be directly obtained from Lemma 1. □

B. ZF Precoding Analysis

Considering now ZF precoding, it follows that the power control factor ρ_ψ is given by $\rho_\psi=\sqrt{P_t/\mathrm{Tr}\{(\mathbf{H}_\psi\mathbf{H}_\psi^H)^{-1}\}}$. Since for ZF all intra-group interference in (4) is completely canceled, the SINR at a typical user $\mathrm{U}_{\psi,k}$ is given by

$$\mathrm{SINR}_{\psi,k}^{\mathrm{ZF}}=P_t\big(G\mathrm{Tr}\big\{\left(\mathbf{H}_\psi\mathbf{H}_\psi^H\right)^{-1}\big\}\big)^{-1}. \quad (11)$$

Now, we can present our next result for the effective sum-rate of ZF-based vector coded caching, which we denote as $\bar{R}_{\mathrm{sum}}^{\mathrm{ZF}}$.

Lemma 2. *For $Q<L$, the effective sum-rate of ZF-based vector coded caching is bounded as $\widetilde{R}_{\mathrm{sum}}^{\mathrm{ZF}}\leq\bar{R}_{\mathrm{sum}}^{\mathrm{ZF}}\leq\widehat{R}_{\mathrm{sum}}^{\mathrm{ZF}}$, where*

$$\widetilde{R}_{\mathrm{sum}}^{\mathrm{ZF}}\triangleq\xi_{GQ}GQ\ln\left(1+\tfrac{P_t}{G}\tfrac{L-Q}{Q}\right) \quad (12)$$

$$\widehat{R}_{\mathrm{sum}}^{\mathrm{ZF}}\triangleq\xi_{GQ}GQ\ln\left(1+\tfrac{P_t}{G}\tfrac{L-Q+1}{Q}\right) \quad (13)$$

[4]Note that both uncoded caching and the cacheless settings enjoy the same transmission rate [11] and they only differ in the amount of data to transmit.

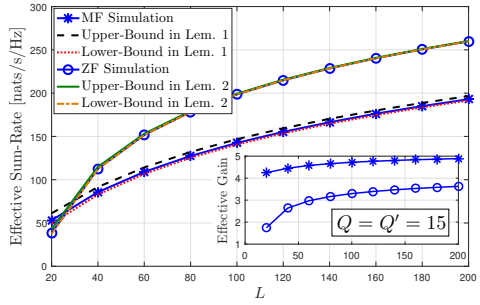

Fig. 1: Effective sum-rate for $P_t = 10$ dB, $G = 6$ and $Q = 15$.

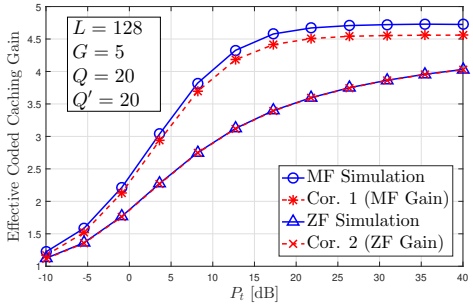

Fig. 2: Effective coded caching gain versus P_t.

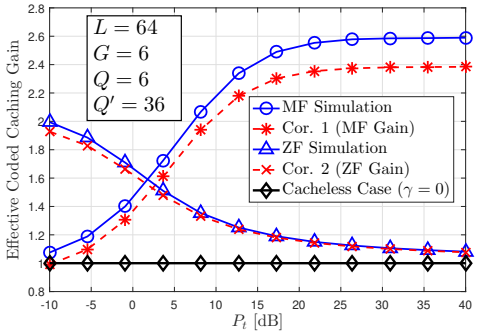

Fig. 3: Effective coded caching gain versus P_t.

Proof. Note that $\mathbf{H}_\psi \mathbf{H}_\psi^H$ is a Wishart matrix with L degrees of freedom. We know from [18] that

$$\mathbb{E}\{\mathrm{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\} = \tfrac{Q}{L-Q}, \text{ for } L > Q. \quad (14)$$

By considering (11) and applying Jensen's inequality on the convex function $\ln(1+x^{-1})$, we can have that

$$\bar{R}_{\mathrm{sum}}^{\mathrm{ZF}} \geq \xi_{GQ} GQ \ln\left(1 + \left(\tfrac{G}{P_t}\mathbb{E}\{\mathrm{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\}\right)^{-1}\right)$$

and thus, by considering (14), we have that (12) lower bounds the performance for this setting.

For the upper-bound, considering that $\mathrm{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\} = \mathrm{Tr}\{\mathbf{V}_\psi^H \mathbf{V}_\psi\} = \sum_{\ell=1}^{Q} ||\mathbf{v}_{\psi,\ell}||^2$, we have that

$$\begin{aligned} & Q\,\mathbb{E}\{\ln\left(1 + \tfrac{P_t}{G}\left(\mathrm{Tr}\left\{\mathbf{H}_\psi^H \mathbf{H}_\psi\right\}\right)^{-1}\right)\} \\ & \overset{(a)}{\leq} \mathbb{E}\Big\{\sum\nolimits_{\ell=1}^{Q} \ln\Big(1 + \tfrac{P_t}{G}\left(Q||\mathbf{v}_{\psi,\ell}||^2\right)^{-1}\Big)\Big\} \\ & \overset{(b)}{\leq} Q \ln\left(1 + \tfrac{P_t}{GQ}\mathbb{E}\left\{\left(||\mathbf{v}_{\psi,k}||^2\right)^{-1}\right\}\right), \end{aligned} \quad (15)$$

where (a) follows from Arithmetic-geometric inequality [5, Lem. 5], and (b) follows from Jensen's inequality and holds for any $k \in [Q]$. By further considering that $\mathbb{E}\{\frac{1}{||\mathbf{v}_{\psi,k}||^2}\} = L - Q + 1$ (cf. [19]) in (15), we can derive $\widehat{R}_{\mathrm{sum}}^{\mathrm{ZF}}$ in (13). □

Corollary 2. *The effective gain of ZF-based (G, Q)-vector coded caching, denoted by $\mathcal{G}_{\mathrm{ZF}}$, is lower-bounded by*

$$\mathcal{G}_{\mathrm{ZF}} \geq \frac{\widetilde{R}_{\mathrm{sum}}^{\mathrm{ZF}}(G,Q)}{\widehat{R}_{\mathrm{sum}}^{\mathrm{ZF}}(1,Q')} = \frac{\xi_{GQ} GQ}{\xi_{Q'} Q'} \frac{\ln\left(1 + \frac{P_t}{G}\frac{L-Q}{Q}\right)}{\ln\left(1 + P_t \frac{L-Q'+1}{Q'}\right)}.$$

Proof. Corollary 2 can be directly obtained from Lemma 2. □

IV. Numerical Results

We provide numerical results to validate the accuracy of the derived expressions. We consider that $T = T_c W_c$, where $T_c = 0.04$ s and $W_c = 300$ kHz, which is suitable e.g. for low-mobility users consuming videos. For simplicity, we set $\Theta = 10$, which is high enough to neglect the impact of CSI estimation noise, such that we assume perfect CSI [15], [16].

Fig. 1 plots the effective sum-rate versus L for a practical SNR value of 10 dB. We can see that the effective sum-rate of MF/ZF precoding grows unboundedly as L increases. As is known for the cacheless setting, MF precoding outperforms the ZF precoding in the very low L region but, after a cutoff point, ZF precoding has a higher effective sum-rate. In contrast, MF has a larger effective gain over the uncoded caching counterpart than ZF. This means that coded caching is more effective for MF because it maintains the same *favorable propagation effect [4]* as for uncoded caching while serving more users at a time.

The effective gain of the considered coded caching scheme is plotted in Fig. 2. As we can see, the effective gain of MF precoding reaches the nominal gain (i.e., the high-SNR nominal gain from [11]) after $P_t = 25$ dB (a realistic value). In contrast, the effective gain of ZF precoding just achieves 85% of the nominal gain when P_t is as high as 40 dB. Fig. 3 shows the effective gain when uncoded caching has the same DoF as coded caching. We observe a 30% additional boost in ZF precoding and a 120% additional boost in MF precoding for $P_t = 10$ dB. The main reason is that coded caching enables each user to remove the interference resulted from $(G-1)Q$ other users, thereby saving much spatial diversity provided by multiple antennas, while the interference in each user is entirely addressed by multiple antennas in uncoded caching.

V. Conclusions

We have analyzed the performance of vector coded caching in the massive MIMO regime, and we have compared its performance for both MF and ZF linear precoding schemes. We have derived simple and robust closed-form expressions for lower and upper bounds of the effective sum-rate, taking into account the CSI acquisition costs, which allows us to investigate the corresponding effective gain over the standard (without coded caching) MIMO systems with the same system parameters. The numerical evaluations show that the derived bounds tightly approximate the actual performance. Furthermore, we have shown how, for a given setting with a given number of transmit antennas and SNR values, incorporating coded caching to the multi-user linear precoding achieves a 300% effective gain for

$P_t \geq 10$ dB for both MF and ZF (cf. Fig. 2). These results show how coded caching can be used in tandem with multi-antenna transmissions while maintaining both coded caching and multiplexing gains in realistic scenarios, and motivates the analysis of other aspects such as the impact of imperfect CSI.

Appendix I: Proof of Lemma 1

The average rate for user $\mathrm{U}_{\psi,k}$ before accounting for CSI costs follows from (8) as $\bar{R}^{\mathrm{MF}}_{\psi,k} = \mathbb{E}\{\ln(1+\mathrm{SINR}^{\mathrm{MF}}_{\psi,k})\}$. Applying Jensen's inequality to the convex function $\ln(1+\frac{1}{x})$ yields

$$\bar{R}^{\mathrm{MF}}_{\psi,k} \geq \ln\left(1+\left(\mathbb{E}\left\{\frac{GQL/P_t}{\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}\right|^2}+\sum_{\vartheta\in[Q]\setminus k}\frac{|\mathbf{h}^T_{\psi,k}\mathbf{h}_{\psi,\vartheta}|^2}{|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2}\right\}\right)^{-1}\right).$$

As $\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}$ follows the Gamma distribution with shape parameter L and scale parameter equal to 1 (cf. [9, Footnote 1]), it follows that $\mathbb{E}\left\{1/|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2\right\} = \frac{1}{(L-1)(L-2)}\ \forall L>2$, from which we obtain $\mathbb{E}\{\frac{GQL}{P_t}/|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2\}$. Next, we have that

$$\begin{aligned}\mathbb{E}\Big\{\sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\frac{|\mathbf{h}^T_{\psi,k}\mathbf{h}_{\psi,\vartheta}|^2}{|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2}\Big\} &= \sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\mathbb{E}\Big\{\frac{\mathrm{Tr}\{\mathbf{h}^T_{\psi,k}\mathbf{h}_{\psi,\vartheta}\mathbf{h}^H_{\psi,\vartheta}\mathbf{h}^*_{\psi,k}\}}{|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2}\Big\}\\ &\overset{(a)}{=} \sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\mathrm{Tr}\Big\{\mathbb{E}\Big\{\frac{\mathbf{h}^*_{\psi,k}\mathbf{h}^T_{\psi,k}}{|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2}\Big\}\Big\} \overset{(b)}{=} \sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\mathbb{E}\Big\{\frac{\mathrm{Tr}\{\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}\}}{|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}|^2}\Big\}\\ &= \sum_{\substack{\vartheta\in[Q]\\ \vartheta\neq k}}\mathbb{E}\Big\{\frac{1}{\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}}\Big\} = \frac{Q-1}{L-1}, \end{aligned} \tag{16}$$

where (a) and (b) follow from exchanging the order of expectation and trace operators and the property $\mathrm{Tr}\{\mathbf{AB}\} = \mathrm{Tr}\{\mathbf{BA}\}$. Finally, combining the results above yields $\widetilde{R}^{\mathrm{MF}}_{\mathrm{sum}}$ in (9).

In the following, we will derive the upper-bound $\widehat{R}^{\mathrm{MF}}_{\mathrm{sum}}$ in (10). First, note that we can write $\bar{R}^{\mathrm{MF}}_{\psi,k}$ as

$$\begin{aligned}\bar{R}^{\mathrm{MF}}_{\psi,k} = \mathbb{E}\Big\{\ln\Big(1+\tfrac{P_t}{GQL}\sum\nolimits_{\ell\in[Q]}\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\ell}\right|^2\Big)\Big\}\\ -\mathbb{E}\Big\{\ln\Big(1+\tfrac{P_t}{GQL}\sum\nolimits_{\vartheta\in[Q]\setminus k}\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\vartheta}\right|^2\Big)\Big\}.\end{aligned} \tag{17}$$

By using Jensen's inequality separately to the concave function $\ln(1+x)$ and to the convex function $\ln(1+x^{-1})$ in (17), we can derive an upper-bound of $\bar{R}^{\mathrm{MF}}_{\psi,k}$ as

$$\begin{aligned}\bar{R}^{\mathrm{MF}}_{\psi,k} \leq \ln\Big(1+\tfrac{P_t}{GQL}\sum\nolimits_{\ell\in[Q]}\mathbb{E}\left\{\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\ell}\right|^2\right\}\Big)\\ -\ln\Big(1+\tfrac{P_t}{GQL}\Big(\mathbb{E}\Big\{\Big(\sum\nolimits_{\vartheta\in[Q]\setminus k}|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\vartheta}|^2\Big)^{-1}\Big\}\Big)^{-1}\Big).\end{aligned} \tag{18}$$

Using the similar manipulations as in (16), we can have that

$$\begin{aligned}\sum_{\ell\in[Q]}\mathbb{E}\left\{\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\ell}\right|^2\right\} &= \mathbb{E}\left\{\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k}\right|^2\right\}+\sum_{\ell\in[Q]\setminus k}\mathbb{E}\left\{\left|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\ell}\right|^2\right\}\\ &\overset{(a)}{=} L^2+L+\sum\nolimits_{\ell\in[Q]\setminus k}\mathrm{Tr}\left\{\mathbb{E}\left\{\mathbf{h}^*_{\psi,\ell}\mathbf{h}^T_{\psi,\ell}\right\}\mathbb{E}\left\{\mathbf{h}^*_{\psi,k}\mathbf{h}^T_{\psi,k}\right\}\right\}\\ &= L^2+L+\sum\nolimits_{\ell\in[Q]\setminus k}\mathrm{Tr}\left\{\mathbf{I}_L\mathbf{I}_L\right\} = L^2+QL,\end{aligned} \tag{19}$$

where (a) follows from the fact that $\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,k} \sim \mathrm{Gamma}(L,1)$ whose mean and variance are both L, and after exchanging the order of expectation and trace operators.

By using iterated expectations, we have that:

$$\begin{aligned}&\mathbb{E}\Big\{\Big(\sum\nolimits_{\vartheta\in[Q]\setminus k}|\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\vartheta}|^2\Big)^{-1}\Big\}\\ &= \mathbb{E}_{\mathbf{h}_{\psi,k}}\mathbb{E}\Big\{\tfrac{1}{||\mathbf{h}_{\psi,k}||^2}\Big(\sum\nolimits_{\vartheta\in[Q]\setminus k}\Big|\tfrac{\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\vartheta}}{||\mathbf{h}_{\psi,k}||}\Big|^2\Big)^{-1}\Big|\mathbf{h}_{\psi,k}\Big\}\\ &\overset{(a)}{=} \mathbb{E}_{\mathbf{h}_{\psi,k}}\Big\{\tfrac{1}{||\mathbf{h}_{\psi,k}||^2}\tfrac{1}{Q-2}\Big\} = \tfrac{1}{(L-1)(Q-2)},\end{aligned} \tag{20}$$

where (a) follows from the fact that given $\mathbf{h}_{\psi,k}$, $X_\vartheta \triangleq |\mathbf{h}^T_{\psi,k}\mathbf{h}^*_{\psi,\vartheta}/||\mathbf{h}_{\psi,k}|||^2$ is an exponential random variable with unit-mean, and hence the summation of $\{X_\vartheta\}_{\vartheta\neq k}$ (i.i.d. given $\mathbf{h}_{\psi,k}$) follows the Gamma distribution with shape parameter $Q-1$ and unit-scale parameter (cf. [9, Footnote 1]). Finally, combining (18)-(20), we derive the upper-bound $\widehat{R}^{\mathrm{MF}}_{\mathrm{sum}}$ in (10). □

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] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[3] 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.*, accepted, doi: 10.1109/TWC.2022.3140895.

[4] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[5] Y.-G. Lim, C.-B. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6827–6842, Dec. 2015.

[6] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[7] 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.

[8] I. Bergel and S. Mohajer, "Cache-aided communications with multiple antennas at finite SNR," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1682–1691, Aug. 2018.

[9] 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.

[10] 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.

[11] 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.

[12] H. Zhao, E. Lampiris, G. Caire, and P. Elia, "Multi-antenna coded caching analysis in finite SNR and finite subpacketization," in *Proc. 25th Int. ITG Workshop on Smart Antennas (WSA)*, Nov. 2021, pp. 433–438.

[13] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," University of Oulu, White Paper, Jun. 2020.

[14] A. Malik, B. Serbetci, E. Parrinello, and P. Elia, "Fundamental limits of stochastic shared-cache networks," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4433–4447, Jul. 2021.

[15] M. Kobayashi, G. Caire, and N. Jindal, "How much training and feedback are needed in MIMO broadcast channels?" in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2008, pp. 2663–2667.

[16] 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.

[17] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Vector coded caching multiplicatively boosts the throughput of realistic downlink systems," 2022. [Online]. Available: https://arxiv.org/abs/2202.07047

[18] 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.

[19] 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.