On the complementary roles of massive MIMO and coded caching for content delivery

Khac-Hoang Ngo, Sheng Yang, Mari Kobayashi, Kaibin Huang

On the Complementary Roles of Massive MIMO and Coded Caching for Content Delivery

Khac-Hoang Ngo*, Sheng Yang*, Mari Kobayashi*, and Kaibin Huang†
* L2S, CentraleSupélec, 91190 Gif-sur-Yvette, France
†Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
khachoang.ngo@supelec.fr, {sheng.yang, mari.kobayashi}@centralesupelec.fr, huangkb@eee.hku.hk

***Abstract*—In this paper, we extend the answer for the question: *Do the gains from massive MIMO and coded caching cumulate?* in a simple setting of downlink MIMO channel with Rayleigh quasi-static fading. Under some practical assumptions, the scalability of massive MIMO and coded caching alone with the number of users does not hold. On the other hand, we show that in this setting, the combination of both provides a scalable solution in most practical scenarios. Therefore, MIMO and coded caching are indeed complementary, especially under an optimal power allocation.**

I. Introduction

In the near future, content delivery is about to take up more than 50% of the mobile traffic. To accommodate the traffic expansion, massive MIMO, using a huge number of antennas at the base station to create a large number of degrees of freedom, is a promising solution to increase substantially the spectral efficiency [1]. If the number of transmit antennas can scale with the number of users K, then the total transmission time for all the K requested files does not increase with K since *simultaneous* transmission can be done in the parallel channels created by precoding (e.g. zero forcing). Another solution is caching, that is, exploiting the on-board memory to prefetch popular contents at (or close to) the end users of the network during off-peak hours so that the traffic during peak hours is significantly reduced. Recently, it has been shown that, with the so-called coded caching, the minimum number of total *multicast* transmissions to satisfy the demand of K users goes to constant when $K \to \infty$ [2]. Instead of sending parallel streams as in MIMO, the single stream (multicast) transmission in coded caching conveys information that is simultaneously useful to a large subset of users. A common perception is that both massive MIMO and coded caching are *potentially* scalable solutions alone with respect to (w.r.t.) the number of users. However, the scalability relies on some ideal assumptions that may not hold in real systems as discussed shortly. Therefore, it is of practical and theoretical interest to address the following question from the engineering perspective: *is it beneficial to use both technologies?*

We first argue that neither of the solutions is indeed scalable in wireless channels under some practical assumptions. The scalability of massive MIMO, w.r.t. the number of users ($K \to \infty$), relies on the *vanishing* error of channel state information at the transmitter's side (CSIT), whereas the scalability of coded caching hinges on a *non-vanishing* multicast rate of the channel.

In [3], we partially answered the above question positively in the particular setting of an i.i.d. quasi-static MISO downlink channel with a multi-antenna base station and K single-antenna receivers. We analyzed the content delivery rates of two schemes and the relative merit of coded caching w.r.t. massive MIMO in various regimes of interest, as the number of users K grows. Results suggest that coded caching shall be preferred to massive MIMO when the per-user power decreases or remains constant, or equivalently when the error variance increases or remains constant, w.r.t. the number of users. Such behavior is expected because it is well known that the gain of massive MIMO vanishes in these cases.

To the best of our knowledge, the work in [3] is the first study that quantifies the relative merit between massive MIMO and coded caching. Among a number of recent works studying coded caching in wireless channels [4], [5], [6], [7], [8], the works [6], [7] also consider MISO broadcast channel. However, these works are conceptually different because their scope is on the interplay between the CSI feedback and coded caching.

In this paper, we still use the same quasi-static channel model and extend the answer by considering the mixed strategy of both massive MIMO and coded caching. We first recall the achievable rates of both schemes derived in [3] for this channel, and propose a scalable solution by combining both schemes with simultaneous multicast and unicast with a power-splitting approach. This combination is first proposed in [9] and then investigated in [10] (and the references therein). We then analyse the asymptotic behaviors of this mixed delivery in large K regime, which suggests that to achieve the maximal rate, the coded multicasting should be allocated the majority of power budget. This is validated by numerical examples.

The remainder of the paper is organized as follows. We describe the system model in Section II and the mixed MIMO transmission in Section III. It is followed by the equivalent sum content delivery rate in the MISO broadcast channel in Section IV. Section V provides the asymptotic analysis of the equivalent content delivery rate and Section VI provides some numerical examples. Finally, we conclude the paper with some discussion in Section VII.

Throughout the paper, we use the following notational conventions. For random quantities, we use upper case non-italic letters, e.g., X, for scalars, upper case letters with bold and

This work was financed in part by Huawei Technologies France SASU.

non-italic fonts, e.g., $\mathbf{V}$, for vectors, and upper case letter with bold and sans serif fonts, e.g., M, for matrices. Deterministic quantities are denoted in a rather conventional way with italic letters, e.g., a scalar x, a vector $\boldsymbol{v}$, and a matrix $\boldsymbol{M}$. Logarithms are in base 2. The Euclidean norm of a vector and a matrix is denoted by $\|\boldsymbol{v}\|$ and $\|\boldsymbol{M}\|$, respectively. The transpose and conjugated transpose of $\boldsymbol{M}$ are $\boldsymbol{M}^{\mathsf{T}}$ and $\boldsymbol{M}^{\mathsf{H}}$, respectively.

II. System Model

In this paper, we consider a MISO downlink channel where a base station with n_{t} transmit antennas communicate with K single-antenna users. The channel $\mathsf{H} \in \mathbb{C}^{K\times n_{\mathrm{t}}}$ is assumed to be a quasi-static fading channel, i.e., remains unchanged during the transmission of a whole coded block. For tractability, we assume that the channel is independent and symmetric across users with i.i.d. Rayleigh fading, i.e., $\mathbf{H}_k \sim \mathcal{CN}(0, \mathbf{I}_{n_{\mathrm{t}}})$, $k = 1, \ldots, K$, with $\mathsf{H} = [\mathbf{H}_1 \ \cdots \ \mathbf{H}_K]^{\mathsf{T}}$. The channel state information (CSI) is assumed to be known perfectly at the receiver side, the transmitter only knows an estimate $\hat{\mathsf{H}}$. Receiver k at time t has the observation

$$\mathrm{Y}_k[t] = \mathbf{H}_k^{\mathsf{T}} \boldsymbol{x}[t] + \mathrm{Z}_k[t], \quad t = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_t \in \mathbb{C}^{n_{\mathrm{t}}\times 1}$ is the input vector at time t, with the average power constraint $\frac{1}{n}\sum_{t=1}^{n} \|\boldsymbol{x}_t\|^2 \le P$; the additive noise $\{\mathrm{Z}_k[t]\}$ is assumed to be spatially and temporally white with normalized variance, i.e., $\mathrm{Z}_k[t] \sim \mathcal{CN}(0,1)$, $k = 1, \ldots, K$. Since the additive noise power is normalized, the transmit power P is identified with the total SNR throughout the paper.

In practice, imperfect CSIT is due to a limited resource for downlink channel training and channel feedback in a FDD system, while it is due to the channel estimation error at the base station and/or imperfect calibration in a TDD system. A common model for the imperfect CSIT is

$$\mathbf{H} = \hat{\mathbf{H}} + \tilde{\mathbf{H}} \tag{2}$$

where $\hat{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are the mutually uncorrelated estimated channel and channel estimation error and have variances $1-\sigma^2$ and σ^2, respectively. Since we assume Rayleigh fading, $\hat{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are independent and circularly symmetric Gaussian distributed.

III. MIMO Mixed Transmission

A. Transmission of mixed common and private information

We consider the transmission of signal carrying both *common* information interested by all the users and a set of *private* information required by each user individually. Given the *common* signal $\mathbf{X}_0$ intended for all users and the *private* signal X_k intended for user k, $k = 1, \ldots, K$, the transmit signal is

$$\mathbf{X} = \mathbf{X}_0 + \sum_{k=1}^{K} \mathbf{W}_k X_k, \tag{3}$$

under the total power constraint P, where we omit the time index for simplicity and $\{\mathbf{W}_k\}$ is the precoder to be defined shortly. Denote the power of common signal as P_0 and the power of private signal X_k as P_k, then $\sum_{k=0}^{K} P_k \le P$. Note that the two extreme cases $P_0 = 0$ and $P_0 = P$ correspond to the transmission of private information only and common information only, respectively. We describe the common signal and private signal in the following.

- **Common signal**: Common signal carries the message to be decoded at each receiver. The maximum common information rate is the minimum of the achievable rate among all users. We assume isotropic signaling, i.e., $\mathbf{X}_0 \sim \mathcal{CN}(0, \frac{P_0}{n_{\mathrm{t}}}\mathbf{I})$.
- **Private signal with zero-forcing precoding**: In order to avoid the inter-user interference, we employ zero-forcing (ZF) precoding scheme, which is commonly used for MISO downlink, such that the private signal for each user should be sent in the null space of the other users' signal space. For ZF to work, we assume that the number of users that can be simultaneously served is smaller than the number of antennas, i.e. $K \le n_{\mathrm{t}}$. This is our assumption whenever ZF is used in the following. Under imperfect CSIT (2), the private signal of user k is precoded in the direction $\mathbf{W}_k$ of unit norm, satisfying the following constraints:

$$\hat{\mathbf{H}}_l^{\mathsf{T}} \mathbf{W}_k = 0, \forall\, l \ne k. \tag{4}$$

The overall precoded private signal is therefore

$$\sum_{k=1}^{K} \mathbf{W}_k \mathrm{X}_k. \tag{5}$$

We use i.i.d. Gaussian signaling for tractability, i.e., $\{\mathrm{X}_k\}$ are i.i.d. $\sim \mathcal{CN}(0, P_k)$.

The received signal at user k is

$$\mathrm{Y}_k = \boldsymbol{H}_k^{\mathsf{T}} \mathbf{X}_0 + \mathrm{G}_k \mathrm{X}_k + \sum_{l\ne k} \tilde{\mathrm{G}}_{k,l} \mathrm{X}_l + \mathrm{Z}_k \tag{6}$$

where

$$\mathrm{G}_k := \mathbf{H}_k^{\mathsf{T}} \mathbf{W}_k \sim \mathcal{CN}(0,1), \tag{7}$$

$$\tilde{\mathrm{G}}_{k,l} := \tilde{\mathbf{H}}_k^{\mathsf{T}} \mathbf{W}_l \sim \mathcal{CN}(0, \sigma^2). \tag{8}$$

Note that the above equivalent channel coefficients are not independent between each other.

The signal-to-noise-ratio (SINR) of the common signal at the receiver k is

$$\mathrm{SINR}_k^{(0)}(\mathsf{H}) := \frac{\frac{P_0}{n_t}\|\mathbf{H}_k\|^2}{1 + |\mathrm{G}_k|^2 P_k + \sum_{l\ne k} |\tilde{\mathrm{G}}_{k,l}|^2 P_l}. \tag{9}$$

For any realization $\mathsf{H} = \boldsymbol{H}$, the rate of common signal is

$$R_0 = \log\Big(1 + \min_k \big\{\mathrm{SINR}_k^{(0)}(\boldsymbol{H})\big\}\Big). \tag{10}$$

Remark III.1. *The rate* (10) *can be regarded as the upper bound on the rate of a quasi-static channel. For this rate to be achievable, it is implicitly assumed that the transmitter is aware of the value of this rate and use the corresponding capacity-achieving channel code.*

To avoid using the outage formulation, we consider the long-term average common throughput

$$\bar{R}_0 = \mathbb{E}\left[\log\Big(1 + \min_k \big\{\mathrm{SINR}_k^{(0)}(\mathsf{H})\big\}\Big)\right]. \tag{11}$$

Note that this is merely a statistical measure and is essentially different from the ergodic rate.

After successfully decoding the common signal, the receiver can then substract this signal and decode its private information from the remaining. Thus, the SINR of the private signal at receiver k is

$$\mathrm{SINR}_k(\mathbf{H}) := \frac{|\mathrm{G}_k|^2 P_k}{1+\sum_{l\neq k}|\tilde{\mathrm{G}}_{k,l}|^2 P_l}, \tag{12}$$

and the long-term average private throughput is

$$\bar{R}_k = \mathbb{E}\left[\log\left(1+\mathrm{SINR}_k(\mathbf{H})\right)\right]. \tag{13}$$

In this system, power allocation is important. By optimally allocating the power $\{P_k\}, k=0,...,K$, we can achieve the maximal transmission rate. This problem will be studied later in this paper.

In the following, we focus on the symmetric power allocation of the private signal:

$$P_k = \frac{P-P_0}{K} =: p, \quad k=1,\ldots,K \tag{14}$$

and the achievable rate of private signal is symmetric too,

$$\bar{R}_k = \bar{R}_{\mathrm{sym}}, \quad k=1,\ldots,K. \tag{15}$$

The SINR of the common and private part can be written in this case, respectively,

$$\mathrm{SINR}_k^{(0)} = \frac{\frac{P_0}{n_t}\|\mathbf{H}_k\|^2}{1+p\left(A_k+(K-1)\sigma^2 B_k\right)}, \tag{16}$$

$$\mathrm{SINR}_k = \frac{A_k}{p^{-1}+(K-1)\sigma^2 B_k} \tag{17}$$

where $A_k := |\mathrm{G}_k|^2$, $B_k := \frac{1}{(K-1)\sigma^2}\sum_{l\neq k}|\tilde{\mathrm{G}}_{k,l}|^2$ with $\mathbb{E}[A_k] = \mathbb{E}[B_k] = 1$. The marginal distribution of SINR_k does not depend on k. In this setting, the power allocation problem is equivalent to power splitting, which is to fraction the common signal power P_0 from total power budget P.

Next, we describe two extreme operating configuration when all the power is allocated to either common signal or private signal. These two cases were studied separately in [3].

B. Transmission of common information only

This is the case when we set $P_0 = P$, or equivalently $p = 0$. The common SINR becomes the common signal-to-noise ratio (SNR)

$$\mathrm{SNR}_k^{\text{mc-(0)}}(\mathbf{H}) := \frac{P}{n_{\mathrm{t}}}\|\mathbf{H}_k\|^2. \tag{18}$$

And the total long-term average throughput is

$$\bar{R}_0^{\mathrm{mc}} = \mathbb{E}\left[\log\left(1+\min_k\left\{\mathrm{SNR}_k^{\text{mc-(0)}}\right\}\right)\right]. \tag{19}$$

C. Transmission of private information only

This is the case when we set $P_0 = 0$, or equivalently, $p = P/K$. The private SINR is still written as

$$\mathrm{SINR}_k^{\mathrm{uc}} = \frac{A_k}{p^{-1}+(K-1)\sigma^2 B_k} \tag{20}$$

where A_k and B_k are defined earlier. The total long-term average throughput of user k is

$$\bar{R}_{\mathrm{sym}}^{\mathrm{uc}} = \bar{R}_k = \mathbb{E}\left[\log\left(1+\mathrm{SINR}_k^{\mathrm{uc}}(\mathbf{H})\right)\right]. \tag{21}$$

Remark III.2. *Consider two MISO downlink systems under the same channel condition and symmetric private power allocation:*

1) **System 1** *transmits private information only with total power P.*
2) **System 2** *uses mixed transmission with total power $P' > P$ and allocates the power P for private signal.*

Then the total long-term average throughput of System 1 is equal to the private long-term average thoughput of System 2. This does not hold for common transmission.

IV. Coded Caching with MIMO Delivery

A. Coded caching

Let us consider the scenario with a content server with N equally popular files of F bits. Each user has a cache of size MF bits, where M denotes the cache size measured in files. Further, each user can prefetch their cache during off-peak hours, prior to the actual request. Then, using coded caching [2], [11] under error-free channel, the number of *multicast* transmissions needed to satisfy K distinct demands from K users, denoted as $T(N,M,K)$ is

$$\begin{cases} \left(1-\frac{M}{N}\right)\frac{1}{1/K+M/N}, & \text{centralized caching} \\ \left(1-\frac{M}{N}\right)\frac{1-\left(1-\frac{M}{N}\right)^K}{M/N}, & \text{decentralized caching} \end{cases} \tag{22}$$

where we assume that $K \leq N$; T is normalized by F, the number of bits to transmit is $T(N,M,K)F$. In the following, we focus on centralized coded caching, the behavior for decentralized caching is essentially the same as it can be readily shown by doing the same exercise. Since T only depends on the normalized memory $m := \frac{M}{N}$, we use the notation $T(m,K)$ whenever confusion is not likely. In the rest of the paper, we assume that $n_{\mathrm{t}} \geq K$.

B. Equivalent content delivery rate

Let us assume that the channel between the content server and the K users is the MIMO channel described in the previous section. We define the equivalent content delivery rate as the number of total demanded information bits (including those already in the cache) that can be delivered per unit of time in average. For instance, when $M = N$, then the equivalent content delivery rate is ∞, since each user can have any content instantly. We consider the following cases:

- **Spatial multiplexing**: sending only private streams to serve different users in parallel. In this case, we try to exploit

the multiplexing gain offered by the MIMO channel. To satisfy the demand of user k, i.e., *complete* the F demanded bits (considering some bits may already be inside the user's cache), we need to send $(1-m)F$ bits, which takes $(1-m)F/\bar{R}_k$ unit of time in average. It follows that the equivalent sum content delivery rate of the system is simply

$$R_{\text{uni-c}} = \frac{K\,\bar{R}^{\text{uc}}_{\text{sym}}(K,P,\sigma^2)}{1-m} \quad \text{bits/second/Hz.} \tag{23}$$

- **Coded caching**: sending only common coded streams to serve all users simultaneously. In this case, we try to exploit the global caching gain offered by the Maddah-Ali Niesen scheme. To satisfy the demand of K users, i.e., *complete* in total KF demanded bits, we need to send $T(m,K)F$ bits, taking $T(m,K)F/\bar{R}^{\text{mc}}_0$ unit of time. It means that the sum content delivery rate of the system is

$$R_{\text{mul-c}} = \frac{K\bar{R}^{\text{mc}}_0(K,P)}{T(m,K)} \quad \text{bits/second/Hz.} \tag{24}$$

- **Mixed delivery**: sending the mixed common and private streams to serve the common request to all user simultaneously, and at the same time serve the individual request to defferent users in parallel. In this case, we try to cumulate the multiplexing gain and the global caching gain. We need to send $(1-m)F$ bits in private streams to each user, which takes $(1-m)F/\bar{R}_{\text{sym}}$ unit of time, and send $T(m,K)F$ bits in common coded streams to every user, which takes $T(m,K)F/\bar{R}_0$ unit of time. It follows that the equivalent sum content delivery rate of the system is

$$R_{\text{mix-c}} = \frac{K\bar{R}_{\text{sym}}(K,P,P_0,\sigma^2)}{1-m} + \frac{K\bar{R}_0(K,P,P_0)}{T(m,K)} \quad \text{bits/second/Hz.} \tag{25}$$

Let $R^{\text{mix}}_{\text{uni-c}} := \frac{K\bar{R}_{\text{sym}}(K,P,P_0,\sigma^2)}{1-m}$ and $R^{\text{mix}}_{\text{mul-c}} := \frac{K\bar{R}_0(K,P,P_0)}{T(m,K)}$. Note that the power splitting P_0/P is critical to $R_{\text{mix-c}}$. We would like to find the optimal power splitting to maximize this equivalent mixed sum rate.

V. Asymptotic behaviors and power splitting

The regime of interest is with a large number of users, i.e., $K \to \infty$. The asymptotic behaviors of $R_{\text{uni-c}}$, $R_{\text{mul-c}}$, and $R_{\text{mix-c}}$ depend on (m,P,σ^2). On top of that, the behaviors of $R_{\text{mix-c}}$ depends also on P_0. In the following, the asymptotic notations O, o, Ω, Θ are w.r.t. K, unless explicitly stated.

A. Asymptotic behaviors of spatial multiplexing $R_{uni\text{-}c}$ and coded caching $R_{mul\text{-}c}$

In [3], we studied the asymptotic behaviors of $R_{\text{uni-c}}$, $R_{\text{mul-c}}$ when K is large in different regimes of total power P. We summarize these results as follows.

1) Power-limited regime: $P = \Theta(1), \sigma^2 = \Theta(1)$

$$R_{\text{uni-c}} = \Theta\big(\frac{1}{1-m}\big), \tag{26}$$

$$R_{\text{mul-c}} = \Theta\left(\frac{1+Km}{1-m}\right). \tag{27}$$

2) Fixed per-user power: $P = \Theta(K), \sigma^2 = \Theta(1)$

$$R_{\text{uni-c}} = \Theta\big(\frac{1}{1-m}\big), \tag{28}$$

$$R_{\text{mul-c}} = \frac{1+Km}{1-m}\log(K) + O(1). \tag{29}$$

3) Increasing per-user power: $p = \Theta(K^\alpha), \sigma^2 = \Theta(p^{-1}) = \Theta(K^{-\alpha})$

$$R_{\text{uni-c}} = \begin{cases} \Theta\big(\frac{1}{1-m}K^\alpha\big), & \text{if } \alpha \le 1 \\ \frac{1}{1-m}(\alpha-1)K\log(K) + O(1), & \text{if } \alpha > 1 \end{cases} \tag{30}$$

$$R_{\text{mul-c}} = \frac{1+Km}{1-m}(\alpha+1)\log(K) + O(1). \tag{31}$$

The proof of the scaling of $R_{\text{uni-c}}$ can be derived by establishing the upper and lower bounds. Details are omitted due to the lack of space. The scaling of $R_{\text{mul-c}}$ can be trivially derived from the following lemma.

Lemma 1. *When* $n_t = 1$, $\mathbb{E}\left[\min_k\{\text{SNR}_k^{mc-(0)}\}\right] = \frac{P}{K}$. *For a fixed total transmit power* P, $\bar{R}_0 = \Theta(1/K)$, *i.e., the multicast rate is* vanishing *with* K *with single transmit antenna. When* n_t *scales at least logarithmically with* K, $\mathbb{E}\left[\min_k\{\text{SNR}_k^{mc-(0)}\}\right] = P\Theta(1)$ *when* K *is large, i.e., the multicast rate is* non-vanishing *when the number of transmit antennas scales up with* K.

This lemma can be proved using the Chernoff bound for the random variable $\frac{\|\mathbf{H}_k\|^2}{n_t}$. Details are omitted due to the lack of space. Lemma 1 shows that a large number of transmit antennas are necessary to achieve non-vanishing multicast rate, which is essential for the scalability of coded caching.

Remark V.1. *We see that with a fixed amount of cache memory* M*, the typical regime in which coded caching is beneficial is when* $\alpha \in (-1, 0]$. *The gain is* $\Theta\left((\alpha+1)\log(K)\right)$. *If either the per-user power blows up with* K *or the total power shrinks with* K*, then coded caching is not useful. Such regimes are however not representative in a wireless communication system.*

B. Asymptotic behaviors of mixed delivery

Remark V.2. *$R_{mix\text{-}c}$ scales at least as the maximum of $R^{mix}_{uni\text{-}c}$ and $R^{mix}_{mul\text{-}c}$.*

Now, we analyse the asymptotic behaviors of $R^{\text{mix}}_{\text{uni-c}}$ and $R^{\text{mix}}_{\text{mul-c}}$. Let the total power scale up with K as $P = \Theta(K^\eta)$ for some $\eta \ge 0$ and the power budget for private signal scale as $P - P_0 = \Theta(K^\beta)$ for some $\beta \le \eta$. Then $P_0 = \Theta(K^\eta)$ and $p = P_k = \Theta(K^{\beta-1}), k = 1, \ldots, K$.

First, we see that for a given η, we have no interest of letting the private power decrease as K, i.e. $\beta < 0$. Thus, we look into the following regimes:

1) Non-increasing per-user total power regime: In this regime, $0 \le \eta \le 1$, and the total power scales with K as $P = \Theta(K^\eta)$. The two extreme cases $\eta = 0$ and $\eta = 1$ correspond to the total power limited regime and fixed per-user total power regime, respectively. The estimation

error is bounded away from zero, i.e. $\sigma^2 = \Theta(1)$. We let $0 \le \beta \le \eta$, then $P_0 = \Theta(K^\eta)$ and $p = \Theta(K^{\beta-1})$. According to (17), $\mathrm{SINR}_k = \Theta(1/K)$ with high probability (*w.h.p.*), and hence $\bar{R}_{\text{sym}} = \Theta(1/K)$ from the linear approximation of $\log(1+x) = x\log e + o(1)$ when $x \to \infty$. From (16) and Lemma 1, $\min_k\{\mathrm{SINR}_k^{(0)}\} = \Theta(K^{\eta-\beta})$ *w.h.p.*, and hence $\bar{R}_0 = \log(1+K^{\eta-\beta}) + O(1)$. Thus, it follows that

$$R_{\text{uni-c}}^{\text{mix}} = \Theta\left(\frac{1}{1-m}\right), \tag{32}$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}\log(1+K^{\eta-\beta}) + O(1). \tag{33}$$

We see that in this regime, the private rate is negligible to the common rate and has no contribution to the scaling of $R_{\text{mix-c}}$. In addition, the scaling of $R_{\text{mix-c}}$ is decreasing with β.

2) Increasing per-user total power regime: In this regime, $\eta > 1$, $P = \Theta(K^\eta)$ and the total power per user P/K scales up with K.

a) Non-increasing per-user private power regime: In this sub-regime, the private power per user is not increasing, i.e. $p = \Theta(K^{\beta-1})$ for $0 \le \beta \le 1$. $P_0 = \Theta(K^\eta)$ and the estimation error is still bounded away from zero, i.e. $\sigma^2 = \Theta(1)$. Then, according to (17), $\mathrm{SINR}_k = \Theta(1/K)$ *w.h.p.* due to the CSIT error, and as in the previous regime $\bar{R}_{\text{sym}} = \Theta(1/K)$. On the other hand, from (16) and applying Lemma 1 to the righthand side numerator, $\min_k\{\mathrm{SINR}_k^{(0)}\} = \Theta(K^{\eta-\beta})$ *w.h.p.*, and hence $\bar{R}_0 = (\eta-\beta)\log(K) + O(1)$. Thus, we have

$$R_{\text{uni-c}}^{\text{mix}} = \Theta\left(\frac{1}{1-m}\right), \tag{34}$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}(\eta-\beta)\log(K) + O(1). \tag{35}$$

Again, we see that the private rate does not contribute to the scaling of the total sum rate. In this sub-regime, the scaling of $R_{\text{mix-c}}$ decreases with β.

b) Increasing per-user private power regime: In this sub-regime, the private power per user can also scale up with K, i.e. $p = \Theta(K^{\beta-1})$ for some $\eta \ge \beta > 1$. The common power is still $P_0 = \Theta(K^\eta)$. We let the estimation error decrease with p as $\sigma^2 = \Theta(p^{-1}) = \Theta(K^{-\beta+1})$, assuming that a training-based scheme is used for channel estimation (see e.g. [12]). In this case, $\mathrm{SINR}_k = \Theta(K^{\beta-2})$ *w.h.p.*. If $\beta < 2$, the SINR still vanishes and $\bar{R}_{\text{sym}} = \Theta(K^{\beta-2})$. If $\beta = 2$ and $\bar{R}_{\text{sym}}$ is also $\Theta(K^{\beta-2})$. If $\beta > 2$, the SINR scales up and $\bar{R}_{\text{sym}}$ becomes logarithmic $(\beta-2)\log(K) + O(1)$. On the other hand, according to (16) and Lemma 1, $\min_k\{\mathrm{SINR}_k^{(0)}\} = \Theta(K^{\eta-(\beta-2)^+-1})$ *w.h.p.*, and $\bar{R}_0 = (\eta-(\beta-2)^+-1)\log(K) + O(1)$ where $(\beta-2)^+ := \max(\beta-2, 0)$. Thus, it follows that

$$R_{\text{uni-c}}^{\text{mix}} = \begin{cases} \Theta\left(\frac{1}{1-m}K^{\beta-1}\right), & \text{if } 1 < \beta \le 2 \\ \frac{1}{1-m}(\beta-2)K\log(K) + O(1), & \text{if } \beta > 2 \end{cases} \tag{36}$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}(\eta-(\beta-2)^+-1)\log(K) + O(1), \tag{37}$$

and in particular if $\beta > 2$

$$R_{\text{mix-c}} = \frac{1}{1-m}\left[(\beta-2)K + (\eta-\beta+1)(1+Km)\right]\log(K) + O(1). \tag{38}$$

In this sub-regime, provided that the per-user cache memory is nonzero and fixed, the private rate is still dominated by the common rate if $1 < \beta \le 2$. The sum rate scaling is independent of β given that $1 < \beta \le 2$ and increasing with β given that $2 < \beta \le \eta$.

The above asymptotic behaviors can also be proved by establishing the upper and lower bounds. Details are omitted in this paper.

C. Optimal power splitting

We are interested in finding the optimal power splitting P_0 from the total power P which maximizes the sum content delivery rate $R_{\text{mix-c}}$. From the scaling of $R_{\text{uni-c}}^{\text{mix}}$ and $R_{\text{mul-c}}^{\text{mix}}$, we can summarize the optimal power splitting in assymptotic regime in the following proposition.

Proposition 1. *Consider a content delivery system with a base station with n_t transmit antennas and N files, and K single-antenna receivers each with normalized cache memory M files, adapting a mixed common and private transmission. Then, the optimal power splitting when $K \to \infty$ is*

$$P_{priv} = \begin{cases} \Theta(1), & \text{if } \eta \in \left[0, 1+\frac{1}{1-m}\right] \\ \Theta(K^\eta), & \text{if } \eta \in \left(1+\frac{1}{1-m}, \infty\right) \end{cases} \tag{39}$$

where P_{priv} is the power budget for the private signals, $m := M/N$ and we assume that the total power $P = \Theta(K^\eta)$, $K \le \min\{N, n_t\}$ and $\sigma^2 = \Theta\left(\frac{K}{P-P_0}\right)$.

In particular, when K is finite, we can write the sum rate in empirical form for an approximation, and then apply the gradient descent method to derive the optimal operating point.

VI. Numerical Results

We show an example to illustrate the equivalent content delivery rate and optimal power splitting with finite (M, N, K, P, σ^2): $N = 2000, n_t = K = 100, P/K = 20$ dB, and $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$. First, in Figure 1, we plot the equivalent sum content delivery rate of mixed delivery as a function of common power fraction P_0/P in different cases of cache memory size M. In general, the sum rate increases with M, and for a fixed $M > 0$, the sum rate achieves its maximum at P_0/P close to 1. This behavior is predicted in Section V-C about optimal power splitting and can also be verified with the asymptotic analysis of $R_{\text{mix-c}}$. In this figure, we also show the optimal operating points computed by gradient descent method, which agree with the sum rate curves.

Next, in Figure 2, we compare the equivalent sum rate of spatial multiplexing, coded multicasting and mixed delivery with optimal power splitting, as a function of cache memory M. We observe that optimal mixed transmission is always optimal in general. For example, we can achieve more than 150% gain

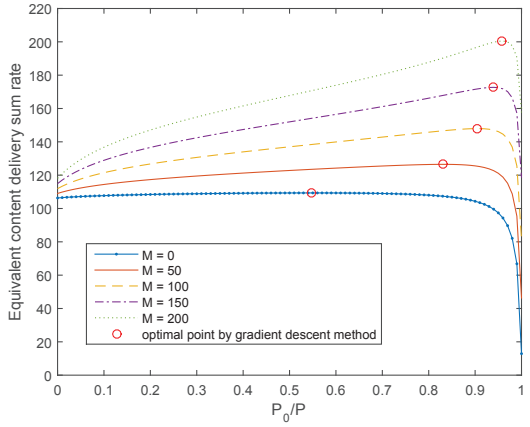

Fig. 1. The equivalent sum content delivery rate as a function of common signal power fraction P_0/P, and the optimal operating point computed by gradient descent method for $N = 2000$, $K = 100$, $P/K = 20$ dB, $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$.

by combining both schemes w.r.t. either one when M is about 140. When M is smaller than that, spatial multiplexing is better than coded multicasting. On the other hand, when M is larger, coded caching is better and is optimal when M is over a certain ratio of library, namely 22%.

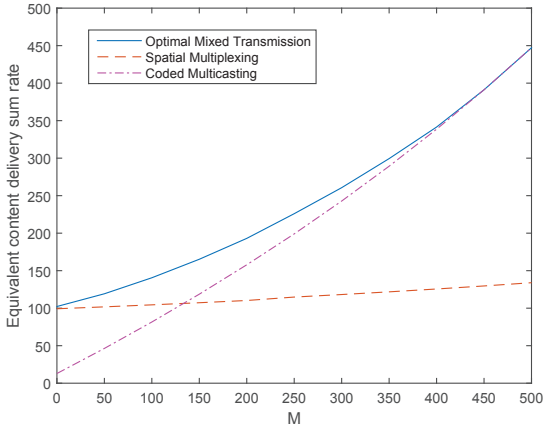

Fig. 2. The equivalent sum content delivery rate of optimal mixed transmission, spatial multiplexing and coded multicasting with user cache as a function of cache memory M for $N = 2000$, $K = 100$, $P/K = 20$ dB, $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$.

Finally, to depict the optimal power splitting, we plot the optimal common power fraction P_0/P, as a function of cache memory M in Figure 3. As M increases, the figure suggests us to allocate more power to the common signal, and even give all the power P to the common signal when M is larger than a certain ratio of the library as named above.

VII. Conclusion

In this paper, we have extended positively the answer for the opening question w.r.t. [3] by showing that massive MIMO and coded caching are indeed complementary to provide a scalable solution for content delivery. They can be combined to improve the equivalent content delivery rate. Several further

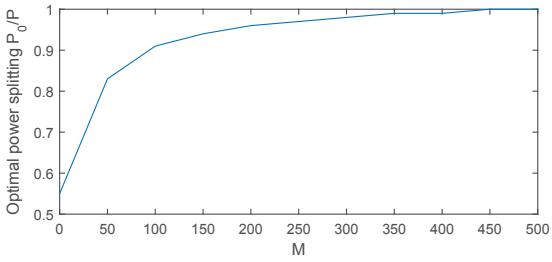

Fig. 3. The optimal power splitting, interpreted by the common power fraction P_0/P, as a function of cache memory M for $N = 2000$, $K = 100$, $P/K = 20$ dB, $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$.

extensions of this work are still open. First, we can consider more practical channel model with high asymmetry according to the user location and antenna correlation, and propose clustering algorithms to maximize the equivalent content delivery rate. Then, coded caching is performed within each cluster but not across different ones. Next, also in the asymmetric channel setting, we can derive the rate regions and provide a utility optimization framework to guarantee the fairness between users.

References

[1] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive mimo for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, 2014.

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

[3] S. Yang, K.-H. Ngo, and M. Kobayashi, "Content delivery with coded caching and massive mimo in 5g," in *9th International Symposium on Turbo Codes & Iterative Information Processing*, Brest, France, Sep. 2016.

[4] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," *arXiv preprint arXiv:1505.01016*, 2015.

[5] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 809–813.

[6] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing csit-feedback in wireless communications," in *Proc. Allerton Conf. Communication, Control and Computing, Monticello, Illinois, USA*, 2015.

[7] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless bc: Interplay of coded-caching and csit feedback," *arXiv preprint arXiv:1511.03961*, 2015.

[8] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," in *the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton),IL, USA,*, 2015.

[9] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated miso broadcast channel with delayed csit," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 315–328, 2013.

[10] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive mimo with imperfect csit," *arXiv preprint arXiv:1512.07221*, 2015.

[11] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015. [Online]. Available: http://dx.doi.org/10.1109/TNET.2014.2317316

[12] 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.