Content delivery with coded caching and massive MIMO in 5G

Sheng Yang, Ngo Khac-Hoang, Mari Kobayashi

Content Delivery with Coded Caching and Massive MIMO in 5G

Sheng Yang, Khac-Hoang Ngo, and Mari Kobayashi
L2S, CentraleSupélec
91190 Gif sur Yvette, France
Email: {sheng.yang, mari.kobayashi}@centralesupelec.fr, khachoang.ngo@supelec.fr

***Abstract*—Do the gains from massive MIMO and coded caching cumulate?** **In this paper, we try to answer the question in a simple setting of downlink MIMO channel with Rayleigh quasi-static fading. While it is generally perceived that each of massive MIMO and coded caching is scalable alone with the number of users, we show that in this setting MIMO and coded caching are indeed complementary and the combination of both provides a scalable solution in most practical scenarios.**

I. Introduction

Content delivery is about to take up more than 70% of the mobile traffic in the near future. 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 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?*

Before trying to answer the question, we shall first argue that neither of the solutions is indeed scalable in wireless channels under some practical assumptions. The scalability of massive MIMO, with respect to 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 our investigation, we consider an i.i.d. quasi-static MISO downlink channel with a multi-antenna base station and K single-antenna receivers. Unfortunately, the underlying conditions for the scalability are hard to be fulfilled in this setting. Then, we show that the combination of both schemes actually provides a scalable solution, which answers the above question positively in this particular setting. Here, we define an equivalent content delivery rate as a unified metric of the throughput performance. We analyze the content delivery rates of two schemes as well as the relative merit of coded caching with respect to massive MIMO in various regimes of interest, as the number of users K grows. The asymptotic analysis validated by numerical examples suggests 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, with respect to the number of users. Such behavior is expected because it is well known that the gain of massive MIMO vanishes in these cases. As a final remark, our work appears to be the first study that quantifies the relative merit between massive MIMO and coded caching to the best of our knowledge. Among a number of recent works studying coded caching in wireless channels [3], [4], [5], [6], [7], the works [5], [6] consider the MISO broadcast channel as the current work. However, these works are conceptually different because their scope is on the interplay between the CSI feedback and coded caching.

The remainder of the paper is organized as follows. The system model is presented in Section II, followed by the unicast rate and the multicast rate in the MISO broadcast channel in Section IV. Section V provides the asymptotic analysis of the unicast/multicast rates 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 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

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

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., remain 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 process $\{\mathrm{Z}_k[t]\}$ 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 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 Transmission

A. Transmission of private information: Zero-forcing precoding

A commonly used precoding algorithm for MISO downlink is the zero-forcing (ZF) precoding scheme such that the 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 signal of user k is precoded in the direction $\mathbf{W}_k$, satisfying the following constraints:

$$\|\mathbf{W}_k\| = 1 \tag{3}$$

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

The precoded signal is therefore

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

where we omit the time index for simplicity. Here X_k is the *private* signal for user k. We use i.i.d. Gaussian signaling for tractability, i.e., $\{\mathrm{X}_k\}$ are i.i.d. $\sim \mathcal{CN}(0, P_k)$, with power constraint $\sum_{k=1}^{K} P_k \le P$.

The received signal at user k is

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

$$= \mathrm{G}_k\mathrm{X}_k + \sum_{l\ne k} \tilde{\mathrm{G}}_{k,l}\mathrm{X}_l + \mathrm{Z}_k \tag{7}$$

where

$$\mathrm{G}_k := \mathbf{H}_k^{\mathsf{T}}\mathbf{W}_k \sim \mathcal{CN}(0,1), \tag{8}$$

$$\tilde{\mathrm{G}}_{k,l} := \tilde{\mathbf{H}}_k^{\mathsf{T}}\mathbf{W}_l \sim \mathcal{CN}(0,\sigma^2). \tag{9}$$

Note that the above equivalent channel coefficients are not independent between each other. The signal-to-interference-and-noise ratio (SINR) at receiver k

$$\mathrm{SINR}_k(\mathsf{H}) := \frac{|\mathrm{G}_k|^2 P_k}{1 + \sum_{l\ne k} |\tilde{\mathrm{G}}_{k,l}|^2 P_l}. \tag{10}$$

Assuming that each user k knows SINR_k, for any realization $\mathsf{H} = \boldsymbol{H}$, we are interested in the following rate

$$R_k(\boldsymbol{H}) = \log\left(1 + \mathrm{SINR}_k(\boldsymbol{H})\right). \tag{11}$$

Remark III.1. *The rate* (11) *can be regarded as an 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 uses the corresponding capacity-achieving channel code.*

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

$$\bar{R}_k = \mathbb{E}\left[\log\left(1 + \mathrm{SINR}_k(\mathsf{H})\right)\right]. \tag{12}$$

Remark III.2. *Note that this is different from the ergodic rate in that the ergodic rate is actually achievable by assuming that a single transmission spans over an infinite number of channel state realizations in an ergodic way (fast fading or block fading model). Here, the average throughput is merely a statistical measure, i.e., the mean of the state-dependent throughput of a quasi-static channel (slow fading model). The two types of models are essentially different. In this paper, we are more interested in the low mobility case and thus the quasi-static channel.*

In the following, we focus on symmetric power allocation,

$$P_k = \frac{P}{K} =: p. \tag{13}$$

Thus, the achievable rate is symmetric too,

$$\bar{R}_k = \bar{R}_{\mathrm{sym}}, \quad k = 1, \ldots, K, \tag{14}$$

$$\sum_k \bar{R}_k = K\bar{R}_{\mathrm{sym}}. \tag{15}$$

The SINR is simplified in this setting

$$\mathrm{SINR}_k = \frac{A_k}{p^{-1} + (K-1)\sigma^2 B_k} \tag{16}$$

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.

B. Transmission of common information

Common information is the message to be decoded at each receiver. The maximum common information rate is the minimum of the achievable rate among all users. Let $\mathrm{X}_0 \sim \mathcal{CN}(0, \boldsymbol{Q}_0)$ be the signal carrying common information, then the common rate is

$$R_0(\boldsymbol{H}) = \max_{\boldsymbol{Q}_0:\mathrm{tr}(\boldsymbol{Q}_0)\leq P} \min_{k\in\{1\dots K\}} \log(1+\boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{h}_k^*). \quad (17)$$

Assuming isotropic signaling, i.e., $\mathrm{X}_0 \sim \mathcal{CN}(0, \frac{P}{n_{\mathrm{t}}}\mathbf{I})$, we have $R_0(\boldsymbol{H}) = \log\left(1+\frac{P}{n_{\mathrm{t}}}\min_k\{\|\boldsymbol{h}_k\|^2\}\right)$. Let us define the common signal-to-noise ratio (SNR) as

$$\mathrm{SNR}_k^{(0)}(\mathbf{H}) := \frac{P}{n_{\mathrm{t}}}\|\boldsymbol{h}_k\|^2. \quad (18)$$

And the long-term average throughput is

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

Lemma 1. *When $n_t = 1$, $\mathbb{E}\left[\min_k\left\{\mathrm{SNR}_k^{(0)}\right\}\right] = \frac{P}{K}$. For a fixed total transmit power P, $\bar{R}_0 = \Theta(1/K)$ when K is large, i.e., the multicast rate is* vanishing *with K with single transmit antenna. When $n_t = K$, $\mathbb{E}\left[\min_k\left\{\mathrm{SNR}_k^{(0)}\right\}\right] = P\Theta(1)$ 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 extreme value theory [8]. Details are omitted due to the lack of space. The above lemma 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.

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], [9], the number of *multicast* transmissions needed to satisfy K distinct demands from K users, denoted as $T(N, M, K)$ is

$$\begin{cases}\left(1-\frac{M}{N}\right)\frac{1}{1/K+M/N}, & \text{centralized caching}\\ \left(1-\frac{M}{N}\right)\frac{1-\left(1-\frac{M}{N}\right)^K}{M/N}, & \text{decentralized caching}\end{cases} \quad (20)$$

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 $n_{\mathrm{t}} = K$.

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 two extreme 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

$$R_{\text{uni-c}} = \frac{K\,\bar{R}_{\text{sym}}(K, P, \sigma^2)}{1-m} \quad \text{bits/second/Hz} \quad (21)$$

  where we write $\bar{R}_{\text{sym}}$ as a function of (K, P, σ^2).
- **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, which takes $T(m, K)F/\bar{R}_0$ unit of time. It means that the sum content delivery rate of the system is simply

$$R_{\text{mul-c}} = \frac{K\bar{R}_0(K, P)}{T(m, K)} \quad \text{bits/second/Hz} \quad (22)$$

  where we write $\bar{R}_0$ as a function of (K, P).

We are particularly interested in the ratio γ between $R_{\text{mul-c}}$ and $R_{\text{uni-c}}$:

$$\gamma = \frac{R_{\text{mul-c}}}{R_{\text{uni-c}}} = \frac{(1-m)K}{T(m, K)}\frac{\bar{R}_0(K, P)}{\bar{R}_{\text{sum}}(K, P, \sigma^2)} \quad (23)$$

$$= (1/K + m)\frac{\bar{R}_0(K, P)}{\bar{R}_{\text{sym}}(K, P, \sigma^2)}. \quad (24)$$

We observe that the gain only depends on the quadruple (m, P, K, σ^2).

V. Large K regime

The regime of interest is the one with a large number of users, i.e., $K \to \infty$. The asymptotic behaviors of $R_{\text{uni-c}}$, $R_{\text{mul-c}}$, and γ depend on (m, P, σ^2). In the following, the asymptotic notations O, o, Ω, Θ are with respect to K, unless explicitly stated.

A. Power-limited regime

In this regime, the total power P is fixed $P = \Theta(1)$, so is the estimation error $\sigma^2 = \Theta(1)$. Then, according to (16), $\mathrm{SINR}_k = \Theta(1/K)$ 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 $x \to 0$. Since $\text{SNR}_k^{(0)} = \Theta(1)$ *w.h.p.*, $\bar{R}_0 = \Theta(1)$. Thus, it follows that

$$R_{\text{uni-c}} = \Theta\big(\frac{1}{1-m}\big), \tag{25}$$

$$R_{\text{mul-c}} = \Theta\left(\frac{1+Km}{1-m}\right), \tag{26}$$

$$\gamma = \Theta\left(1+Km\right). \tag{27}$$

We see that in this regime, unless the cache memory per user vanishes with the number of users as $m = O(1/K)$, coded caching is beneficial.

B. Fixed per-user power p

In this regime, the total power P is increasing with $P = \Theta(K)$ in such a way that for any K each user receives a constant amount of power p when perfect ZF is applied. The CSI estimation error is still bounded away from zero, i.e., $\sigma^2 = \Theta(1)$). Then, according to (16), $\text{SINR}_k = \Theta(1/K)$ *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, since the total power is increasing with K, $\text{SNR}_k^{(0)} = \Theta(P) = \Theta(K)$ *w.h.p.*, thus $\bar{R}_0 = \log(K) + O(1)$. We have

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

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

$$\gamma = \Theta\big((1+Km)\log(K)\big). \tag{30}$$

We see that in this regime, the gain is increasing with K even without cache ($m = 0$). Essentially, this is because multicast rate can scale up with K as $\log(K)$ regardless of CSIT error, while the unicast rate is limited by the interference caused by imperfect CSIT.

C. Increasing per-user power

In this regime, the per-user power can also scale up with K, i.e., $p = \Theta(K^\eta)$ for some $\eta > 0$. We also let the estimation error decrease with p as $\sigma^2 = \Theta(p^{-1}) = \Theta(K^{-\eta})$, assuming that a training-based scheme is used for channel estimation (see e.g. [10]). In this case, $\text{SINR}_k = \Theta(K^{\eta-1})$ *w.h.p.*. If $\eta < 1$, the SINR still vanishes and $\bar{R}_{\text{sym}} = \Theta(K^{\eta-1})$. If $\eta = 1$, the SINR is $\Theta(1)$ and thus $\bar{R}_{\text{sym}}$ is also $\Theta(K^{\eta-1})$. If $\eta > 1$, the SINR scales up and $\bar{R}_{\text{sym}}$ becomes logarithmic $(\eta-1)\log(K)+O(1)$. For multicast, $\text{SNR}_k^{(0)} = \Theta(P) = \Theta(K^{\eta+1})$ *w.h.p.*, and $\bar{R}_0 = (\eta+1)\log(K) + O(1)$. Thus, it follows that

$$R_{\text{uni-c}} = \begin{cases} \Theta\big(\frac{1}{1-m}K^\eta\big), & \text{if } \eta \le 1 \\ \frac{1}{1-m}(\eta-1)K\log(K) + O(1), & \text{if } \eta > 1 \end{cases} \tag{31}$$

$$R_{\text{mul-c}} = \frac{1+Km}{1-m}(\eta+1)\log(K) + O(1) \tag{32}$$

$$\gamma = \begin{cases} \Theta\big((1+Km)(\eta+1)K^{-\eta}\log(K)\big), & \text{if } \eta \le 1 \\ \frac{1+Km}{K}\frac{\eta+1}{\eta-1} + o(1), & \text{if } \eta > 1 \end{cases} \tag{33}$$

In this regime, the per-user cache memory plays an important role. Without cache memory ($m = 0$), unicast is always better than multicast with coded caching as long as the per-user power scales up ($\eta > 0$), i.e., $\gamma = O(1)$. With a constant $m > 0$, the situation is reversed, i.e., the gain γ is $\Omega(1)$ for any $\eta > 0$.

We summarize the asymptotic behavior of the caching gain in the following proposition. Details are omitted due to the lack of space.

Proposition 1. *Consider a content delivery system with a base station (content server) with K transmit antennas and N files, and K single-antenna receivers each with cache memory M files. Then, the coded caching gain γ in terms of the equivalent content delivery rate, when $K \to \infty$, is*

$$\gamma = \begin{cases} \Theta\left((1+Km)P\right), & \text{if } \eta \in (-\infty, -1] \\ \Theta\left((1+Km)\frac{\log(1+P)}{1+P/K}\right), & \text{if } \eta \in (-1, 1] \\ \frac{1+Km}{K}\frac{\eta+1}{\eta-1} + o(1), & \text{if } \eta \in (1, \infty) \end{cases} \tag{34}$$

where we assume that the per-user power $P/K = \Theta(K^\eta)$; $m := M/N$ with $N \ge K$.

Corollary 1. *With a fixed amount of cache memory M, the coded caching gain becomes*

$$\gamma = \begin{cases} \Theta\left(P\right), & \text{if } \eta \in (-\infty, -1], \\ \frac{\log(1+P)}{1+P/K}, & \text{if } \eta \in (-1, 1], \\ \Theta\left(\frac{\eta+1}{\eta-1}\frac{1}{K}\right), & \text{if } \eta \in (1, \infty). \end{cases} \tag{35}$$

Remark V.1. *From the above corollary, we see that with a fixed amount of cache memory M, the typical regime in which coded caching is beneficial is when $\eta \in (-1, 0]$. The gain is $\Theta\left((\eta+1)\log(K)\right)$. 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.*

Remark V.2. *Although we have assumed imperfect CSIT so far, we can extend our analysis to the case with perfect CSIT and show that coded caching is never beneficial compared to massive MIMO. Namely, the relative merit of coded caching holds only when the normalized cache size m grows as $\Omega(K^\eta)$ for $\eta > 0$. This is impossible under the assumption $m \le 1$.*

VI. Numerical Results

We show an example to illustrate the gain γ with finite (M, N, K, P, σ^2). First, we calculate the equivalent sum content delivery rate of the system as a function of per-user cache memory size M for $N = 2000, K = 100$, in different cases of per-user power $p = P/K$. We consider $p = 10, 20, 30, 40$ dB. In each case, we let the CSIT error, when it is present, $\sigma^2 = 1/p$ to avoid the excessive number of parameters. The sum rate of the system with spatial multiplexing $R_{\text{uni-c}}$ (21) and with coded caching $R_{\text{mul-c}}$ (22) are shown in Figure 1 and Figure 2, respectively. We observe that while spatial multiplexing suffers from CSIT estimation error, coded caching is robust under inperfect CSIT. Then, in Figure 3, we plot the coded caching gain γ (24) as a function of per-user cache memory size M in

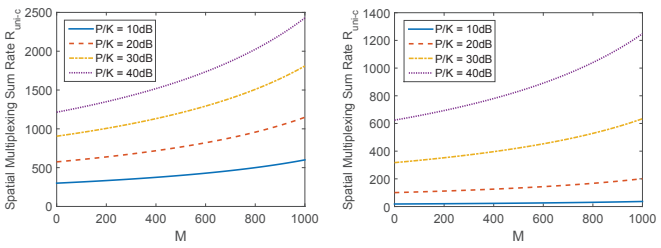

Fig. 1. Spatial multiplexing sum content delivery rate for $N = 2000$, $K = 100$: (a) perfect CSIT, (b) imperfect CSIT.

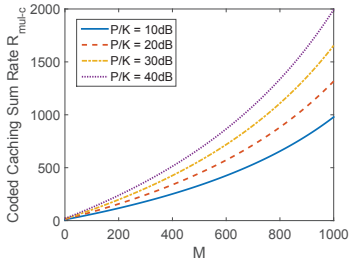

Fig. 2. Coded caching sum content delivery rate for $N = 2000$, $K = 100$. The result remains for perfect CSIT and imperfect CSIT.

the same setting. Comparing the curves, we confirm that coded caching can be quite beneficial especially when the CSIT error is taken into account. For example, for a cache size of 300 files (15% of the library) and $p = 10$ dB, coded caching is twice worse than spatial multiplexing under perfect CSIT, while it is 8.4 times better than spatial multiplexing under imperfect CSIT. As we can see, when M is large enough, the coded caching gain grows linearly with M. This can also be verified with the asymptotic analysis in Proposition 1. The slope of this linear behavior is closely related to the per-user power P/K, as we can see in Proposition 1.

VII. Discussions

Until now, we have considered a rather simple setting in which we investigated the benefit of coded caching in terms of the equivalent content delivery rate. Several extensions are foreseen in the future. First, the results rely on the fact that $n_t = K$ or $n_t = \Theta(K)$. It is sometimes important to measure the impact of the number of transmit antennas. We may use extreme value theory [8] to include n_t in the analysis. Then, instead of doing multicast or spatial multiplexing alone, we can combine both to perform simultaneous multicast and unicast, as it has been first proposed in [11] and then investigated in [12] (and the references therein), to optimize the equivalent content delivery rate. The synergy of coded caching with multicast and spatial multiplexing can bring a substantial performance gain. Last, we have considered symmetric downlink channel in this work. In practice, the channel can be highly asymmetric according to the location of the users and antenna correlation. In this case, taking these factors into account (as modeled in [13]), we can propose clustering algorithms to maximize the equivalent

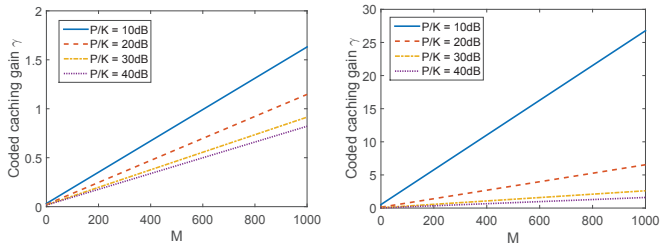

Fig. 3. Coded caching gain over spatial multiplexing for $N = 2000$, $K = 100$; (a) perfect CSIT, (b) imperfect CSIT.

content delivery rate. Then, coded caching is performed within each cluster but not across different ones.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] H. A. David and H. N. Nagaraja, *Order statistics*. Wiley Online Library, 1970.

[9] 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

[10] 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.

[11] 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.

[12] 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.

[13] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing – the large-scale array regime," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6441–6463, 2013.