Cache-aided content delivery in MIMO channels

Khac-Hoang Ngo, Sheng Yang, Mari Kobayashi

Cache-Aided Content Delivery in MIMO Channels

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

***Abstract*—In this paper, we investigate the content delivery problem in the context of multi-antenna (MIMO) wireless networks. The single-antenna users, equipped with some cache memory, receive requested contents from the server through a multi-antenna base station. We propose a scheme that carefully combines the multicast and unicast capabilities offered by MIMO, as a function of the quality of channel state information at the transmitter side. Thereby we reveal the complementary roles of coded caching and MIMO transmission for content delivery.**

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 (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?*

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, w.r.t. the number of users ($K \to \infty$), hinges on: 1) the linearly increasing number of the transmit antennas with respect to the number of users, and 2) the accuracy of channel state information at the transmitter's side (CSIT). The scalability of coded caching depends on a *non-vanishing* multicast rate of the channel. In this work, we consider the specific model of quasi-static i.i.d. (independent and identically distributed) Rayleigh fading downlink channel with a multi-antenna base station and K single-antenna receivers. The quasi-static assumption may be justified, e.g., in low mobility scenario or the latency constrained applications such as the video streaming with independently coded/decoded chunks. In this particular setting, we first show that the combination of coded caching and MIMO multicasting is a scalable solution even when the above constraints are not satisfied. This is due to the channel hardening effect by increasing the number of transmit antennas. In this sense, we answer the above question positively: coded caching and MIMO complement each other. Then, we propose a scheme that exploits the spatial multiplexing capability of MIMO in addition to the channel hardening effect.

To the best of our knowledge, our work appears to be 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 [3], [4], [5], [6], [7], the works [5], [6] consider the fading 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 combination of multicast and spatial multiplexing in the presence of CSIT error was first proposed in [8] and then investigated in [9] (and the references therein).

The rest of the paper is organized as follows. The channel model is described in Section II. Some basic results on spatial multiplexing and multicast in MIMO channels are presented in Section III. We introduce coded caching with MIMO delivery in Section IV. We propose the simultaneous multiplexing and multicasting with coded caching in Section V. Numerical results are shown in Section VI followed by some conclusions in Section VII. Proof details can be found in the appendix.

Throughout the paper, we use the following notational conventions. $\mathrm{X}, \mathbf{V}, \mathbf{M}$ are used to denote random scalars, vectors, matrices, while $x, \boldsymbol{v}, \boldsymbol{M}$ are used to denote deterministic scalars, vectors, matrices. 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. The dot-equality $a \doteq K^b$ means $\lim_{K\to\infty} \frac{\log a}{\log K} = b$, while the dot-inequality is defined similarly. The asymptotic notations O, o, Ω, Θ are w.r.t. to K, unless explicitly stated.

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

II. Channel Model

In this paper, we consider a multi-antenna downlink channel where a base station with n_{t} transmit antennas communicates 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 $\mathsf{H} = [\mathbf{H}_1 \ \cdots \ \mathbf{H}_K]^{\mathsf{T}}$. 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 additive noise power is normalized, the transmit power P is identified with the total signal-to-noise ratio (SNR) throughout the paper.

In practice, the channel state information (CSI) is not perfectly known at the transmitter, typically due to limited resource for uplink channel training in TDD (time division duplex) or limited channel feedback bandwidth in FDD (frequency division duplex). A common model for the imperfect CSIT, modeling the MMSE channel estimation, is

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

where $\hat{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are the mutually uncorrelated estimate and estimation error, of variance $1-\sigma^2$ and σ^2, respectively. Since we assume Rayleigh fading, $\hat{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are independent and circularly symmetric Gaussian distributed. We assume that CSI is perfect at the receivers.

III. MIMO: Spatial Multiplexing vs. Multicasting

In the following, we review the two different uses of MIMO in a downlink channel.

A. Spatial multiplexing

The goal is to create K parallel channels to individual users in such a way that they can communicate with the base station simultaneously with an acceptable rate. Spatial multiplexing relies on precoding: steer the signal to the desirable direction according to the available CSIT. The transmitted signal is

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

where X_k is the *private* signal for user k and $\mathbf{W}_k$ is the precoder for user k of unit norm. Here, we omit the time index for simplicity. By focusing on the zero-forcing (ZF) precoder due to its simplicity for $K \le n_t$, we let $\{\mathbf{W}_k\}$ to satisfy

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

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

$$\mathrm{Y}_k = \mathrm{G}_k \mathrm{X}_k + \sum_{l\ne k} \tilde{\mathrm{G}}_{k,l} \mathrm{X}_l + \mathrm{Z}_k \tag{5}$$

where

$$\mathrm{G}_k := \mathbf{H}_k^{\mathsf{T}} \mathbf{W}_k \sim \mathcal{CN}(0,1), \tag{6}$$

$$\tilde{\mathrm{G}}_{k,l} := \tilde{\mathbf{H}}_k^{\mathsf{T}} \mathbf{W}_l \sim \mathcal{CN}(0, \sigma^2). \tag{7}$$

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

$$\mathrm{SINR}_k(\mathbf{H}) := \frac{|\mathrm{G}_k|^2 P_k}{1 + \sum_{l\ne k} |\tilde{\mathrm{G}}_{k,l}|^2 P_l}. \tag{8}$$

For any realization $\mathsf{H} = \boldsymbol{H}$, we obtain the rate

$$R_k(\boldsymbol{H}) = \log\left(1 + \mathrm{SINR}_k(\boldsymbol{H})\right) \tag{9}$$

which is achievable with suitable rate adaptation and capacty-achieving channel code. Then, the long-term average throughput of the user k is

$$\bar{R}_k := \mathbb{E}\left[\log\left(1 + \mathrm{SINR}_k(\mathbf{H})\right)\right]. \tag{10}$$

One of the important operating points is the symmetric rate with uniform power allocation $P_k = P/K =: p, \forall k$, given by

$$\bar{R}_{\mathrm{sym}} = \bar{R}_k, \quad \forall k. \tag{11}$$

Lemma 1. *In the large K regime, let the per-user power $p := \frac{P}{K} = \Theta(K^{\eta})$ for some $\eta \ge -1$, and $\sigma^2 = \Theta(\min\{p^{-1}, 1\}) = \Theta(\min\{K^{-\eta}, 1\})$. Then, the symmetric rate has the following polynomial behavior*

$$\bar{R}_{sym} \doteq K^{(-1+\eta^+)^-}. \tag{12}$$

where $A^+ := \max\{A, 0\}$ and $A^- := \min\{A, 0\}$.

Proof. See Appendix A. □

Note that in the above lemma, for simplicity, only the polynomial behavior w.r.t. K is shown. For example, the polynomial behavior of $\log(K)$ is $\Theta(1)$. A more refined analysis is done in Appendix A. We remark that, due to the CSIT error that is inversely proportional to the per-user power, the per-user rate is not vanishing only when $\eta \ge 1$, i.e., $P \dot{\ge} K^2$.

B. MIMO Multicasting

The goal of multicasting is to convey a *common* message at the maximum rate so that *every* user can decode. In this case, the message is coded in $\mathbf{X} = \mathbf{X}_0$. Using Gaussian signaling, i.e., $\mathbf{X}_0 \sim \mathcal{CN}(0, \boldsymbol{Q}_0)$, then the common rate is

$$R_0(\boldsymbol{H}) = \max_{\boldsymbol{Q}_0 : \mathrm{tr}(\boldsymbol{Q}_0) \le P} \ \min_{k\in\{1\ldots K\}} \log(1 + \boldsymbol{h}_k^{\mathsf{T}} \boldsymbol{Q}_0 \boldsymbol{h}_k^*). \tag{13}$$

For simplicity, we assume isotropic signaling, i.e., $\mathbf{X}_0 \sim \mathcal{CN}(0, \frac{P}{n_{\mathrm{t}}}\mathbf{I})$, 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 SNR at user k as

$$\mathrm{SNR}_k^{(0)}(\mathbf{H}) := \frac{P}{n_{\mathrm{t}}} \|\mathbf{H}_k\|^2. \tag{14}$$

Then, the long-term multicast thoughput is

$$\bar{R}_0 = \mathbb{E}\left[\log\left(1+\min_k\left\{\mathrm{SNR}_k^{(0)}\right\}\right)\right]. \tag{15}$$

Lemma 2. *When $n_t = 1$, $\min_k\{\mathrm{SNR}_k^{(0)}\}$ is exponentially distributed with mean $\frac{P}{K}$. When n_t scales at least logarithmically with K, i.e. $n_t = \Omega(\log(K))$,*

$$\lim_{K\to\infty}\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \in [1-\epsilon_0, 1+\epsilon]\right) = 1 \tag{16}$$

with $\epsilon_0 \simeq 0.8414$ and any arbitrarily small $\epsilon > 0$.

Proof. See Appendix B. □

This lemma says that the multicasting rate, dominated by the worst user, improves with a sufficiently large number of antennas thanks to channel hardening effect. That is, the fluctuation between K users' channels vanishes and approaches to a constant in the regime of a large K. The following lemma characterizes the contrasted behaviors for a fixed or increasing number of transmit antennas.

Lemma 3. *In the large K regime, let the per-user power $p := \frac{P}{K} = \Theta(K^\eta)$ for some η. Then, the multicast rate scales with K as*

$$\bar{R}_0 \doteq \begin{cases} K^{\eta^-}, & \text{if } n_t = O(1), \\ K^{(\eta+1)^-}, & \text{if } n_t = \Omega(\log(K)). \end{cases} \tag{17}$$

Proof. See Appendix C. □

Again, we focus on the polynomial scaling. We see that with a fixed number of transmit antennas, the multicast rate is vanishing unless the total transmit power is increasing with K such that $P \dot{\geq} K$ (or $\eta > 0$). For an increasing $n_t = \Omega(\log(K))$, the multicasting rate grows under a relaxed condition $\eta > -1$. If n_t grows even faster with K as $n_t = \Omega(K)$, then we can show that a constant transmit power is enough to guarantee the non-vanishing multicast rate.

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], [10] 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

$$\begin{cases} \left(1-\frac{M}{N}\right)\frac{1}{1/K+M/N}, & \text{centralized caching} \\ \left(1-\frac{M}{N}\right)\frac{1-\left(1-\frac{M}{N}\right)^K}{M/N}, & \text{decentralized caching} \end{cases} \tag{18}$$

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. The following lemma is straightforward from the (18).

Lemma 4. *In the large K regime, let $m = \Theta(K^{-\mu})$ for some $\mu > 0$. Then, we have*

$$T(m, K) \doteq K^{\min\{\mu,1\}}. \tag{19}$$

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, 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

$$R_{\text{uni-c}} = \frac{K\,\bar{R}_{\text{sym}}(K, P, \sigma^2)}{1-m} \quad \text{bits/second/Hz} \tag{20}$$

  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

$$R_{\text{mul-c}} = \frac{K\bar{R}_0(K, P)}{T(m, K)} \quad \text{bits/second/Hz} \tag{21}$$

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

The asymptotic behaviors of $R_{\text{uni-c}}$ and $R_{\text{mul-c}}$ are provided in the Appendix D. From (12), (17), and the definitions in (20) and (21), the following proposition follows readily.

Proposition 1. *In the large K regime, let $m = \Theta(K^{-\mu})$ for some $\mu > 0$, and the per-user power $p := \frac{P}{K} = \Theta(K^\eta)$ for some η. Then, we have*

$$R_{\textit{uni-c}} \doteq K^{\min\{\eta,1\}^+}, \tag{22}$$

$$R_{\textit{mul-c}} \doteq K^{(\eta+1)^- + (1-\mu)^+}. \tag{23}$$

As a result, $R_{\textit{mul-c}} \dot{\geq} R_{\textit{uni-c}}$ if and only if $\eta \leq (1-\mu)^+$.

Intuitively, coded caching is beneficial when the per-user power does not scale too fast as compared to the scaling of the memory.

V. Simultaneous Multiplexing and Multicasting

So far, we have shown that MIMO can either be used for spatial multiplexing (i.e. unicast), or for multicast combined with coded caching. The former performs better at high SNR and with precise CSIT, whereas the latter is preferable otherwise. Then, it is natural to combine both the benefits of spatial multiplexing and channel hardening of MIMO transmission. This can be achieved with rate splitting as described as follows. We consider the transmission of signal carrying both common information interested by all the users and a set of private information intended exclusively for each user. Given the *common* signal $\mathbf{X}_0$ dedicated to every user and the *private* signal X_k to user k, $\forall k$, the transmitted signal is

$$\mathbf{X} = \mathbf{X}_0 + \sum_{k=1}^{K} \mathbf{W}_k X_k, \tag{24}$$

where $\mathbf{X}_0, \mathrm{X}_k$, and $\mathbf{W}_k$, $k = 1, \ldots, K$, are defined as before, except for the new total power constraint $\sum_{k=0}^{K} P_k \leq P$. Obviously, this general setting includes the two extreme cases $P_0 = 0$ for spatial multiplexing and $P_0 = P$ for multicast. The received signal at user k is

$$\mathrm{Y}_k = \mathbf{H}_k^{\mathrm{T}} \mathbf{X}_0 + \mathrm{G}_k \mathrm{X}_k + \sum_{l \neq k} \tilde{\mathrm{G}}_{k,l} \mathrm{X}_l + \mathrm{Z}_k \tag{25}$$

where G_k and $\tilde{\mathrm{G}}_{k,l}$ are defined as in (6) and (7).

Each receiver is interested in decoding the common message and its own private message. We consider successive decoding so that each user decodes the common message first and then the private message. Therefore, the private messages are seen as interference to the common message with SINR,

$$\mathrm{SINR}_k^{(0)}(\mathbf{H}) := \frac{\frac{P_0}{n_t} \|\mathbf{H}_k\|^2}{1 + |\mathrm{G}_k|^2 P_k + \sum_{l \neq k} |\tilde{\mathrm{G}}_{k,l}|^2 P_l}, \tag{26}$$

at receiver k, whereas the private messages are decoded as before after removing the decoded common message, with the same SINR as defined in (8). Then, it follows that the equivalent content delivery rate is

$$R_{\text{mix}} = \frac{K \bar{R}_{\text{sym}}(K, P - P_0, \sigma^2)}{1 - m} + \frac{K \bar{R}_0^{\text{mix}}(K, P, P_0)}{T(m, K)} \tag{27}$$

where

$$\bar{R}_0^{\text{mix}}(K, P, P_0) := \mathbb{E}\left[\log\big(1 + \min_k \{\mathrm{SINR}_k^{(0)}\}\big)\right] \tag{28}$$

and we assume symmetric private power allocation. Let $R_{\text{uni-c}}^{\text{mix}} := \frac{K \bar{R}_{\text{sym}}(K, P - P_0, \sigma^2)}{1 - m}$ and $R_{\text{mul-c}}^{\text{mix}} := \frac{K \bar{R}_0^{\text{mix}}(K, P, P_0)}{T(m, K)}$.

The splitting of common power P_0 and private power $P - P_0$ is to be optimized to achieve a maximum delivery rate R_{mix}. If we are only interested in the polynomial behavior of R_{mix} w.r.t. K, then we can easily verify that the exponent is the same as that of $\max\{R_{\text{mul-c}}^{\text{mix}}, R_{\text{uni-c}}^{\text{mix}}\}$. To see this, we remark that $2 \max\{R_{\text{mul-c}}^{\text{mix}}, R_{\text{uni-c}}^{\text{mix}}\} \geq R_{\text{mix}}^{*} \geq \max\{R_{\text{mul-c}}^{\text{mix}}, R_{\text{uni-c}}^{\text{mix}}\}$, both bounds have the same exponent of K. As we shall show in the next section, the performance gain of the simultaneous transmission is considerable with finite K.

Proposition 2. *In the large K regime, let the CSIT error scale as $\sigma^2 = \Theta\left(\frac{K}{P - P_0}\right)$, and the total power scale as $P = \Theta(K^{\eta+1})$. Then, to achieve the optimal scaling of R_{mix}, the total power of private signal is*

$$P - P_0 = \begin{cases} \Theta(1), & \text{if } \eta \in \left[-1, \frac{1}{1-m}\right], \\ \Theta(K^{\eta+1}), & \text{if } \eta \in \left(\frac{1}{1-m}, \infty\right) \end{cases} \tag{29}$$

where $m := M/N$.

Proof. See Appendix E. □

VI. Numerical Results

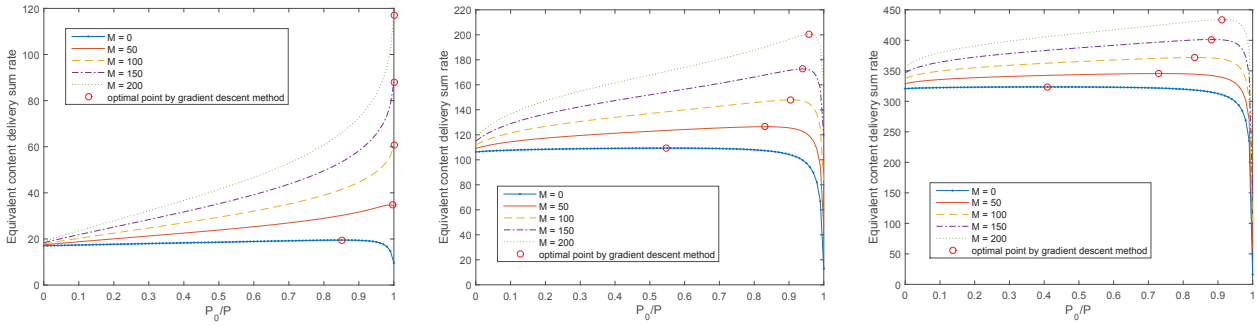
We show an example to illustrate the equivalent sum content delivery rate and optimal power splitting with finite (M, N, K, P, σ^2). The setting is $N = 2000, K = n_{\mathrm{t}} = 100$, and $\sigma^2 = (P/K)^{-1}$ for different cases of per-user total power P/K, namely, 10 dB, 20 dB and 30 dB. First, in Figure 1, we plot the equivalent sum content delivery rate of mixed transmission, i.e., simultaneous multiplexing and multicasting, as a function of common signal power fraction P_0/P in different cases of per-user total power P/K and cache memory size M. In general, for a fixed P/K, the sum rate increases with M, and for a fixed M, the sum rate achieves its maximum at P_0/P close to 1, especially when P/K is not large. This behavior is predicted in the Proposition 2. In this figure, we also show the optimal operating points computed by numerical gradient descent method, which agree with the sum rate curves.

Next, in Figure 2, we compare the equivalent sum rate of mixed transmission under optimal power splitting with spatial multiplexing and coded multicasting alone as a function of cache memory M in different cases of P/K. We observe that optimal mixed transmission is always optimal in general. For example, we can achieve more than 150% gain by combining both schemes w.r.t. either one when M is about 140 and $P/K = 20$ dB. When M is small, spatial multiplexing is better than coded multicasting. On the other hand, when M is large, coded multicasting is better and is optimal when M is larger than a certain ratio of the library, namely, 10%, 22% and 90% for $P/K = 10, 20, 30$ dB, respectively.

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 for different values of P/K. As M increases, the figure suggests us to allocate to the common signal more power, and even all the power when M is larger than a certain fraction of the library as named above.

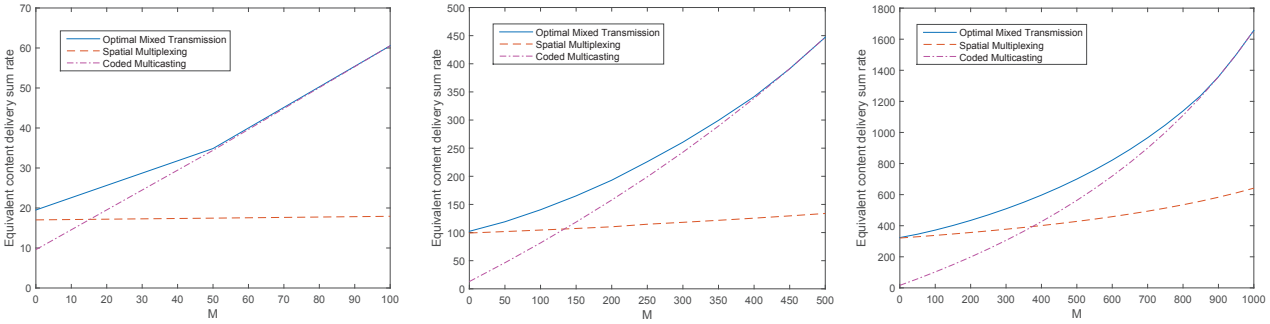
VII. Conclusion

In this paper, we have shown that multiple-antenna transmission is complementary to coded caching to provide a scalable solution for content delivery with a large number of users. Coded caching relieves some practical constraints on MIMO downlink such as linearly increasing number of transmit antennas and accurate channel state information at the transmitter. We have also shown that multiplexing and multicasting can be combined to improve the equivalent



(a) $P/K = 10$ dB (b) $P/K = 20$ dB (c) $P/K = 30$ dB

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



(a) $P/K = 10$ dB (b) $P/K = 20$ dB (c) $P/K = 30$ dB

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 $N = 2000$, $K = 100$, $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$.

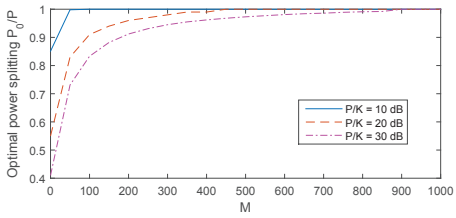


Fig. 3. The optimal power splitting, interpreted by the common power fraction P_0/P, as a function of cache memory M for $N = 2000$, $K = 100$, $P/K = 10, 20, 30$ dB, $\sigma^2 = \left(\frac{P}{K}\right)^{-1}$.

content delivery rate. On-going works include establishing upper bounds on the delivery rate and considering spatially correlated 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] 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] 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.

[9] 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.

[10] 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

APPENDIX

A. Proof of Lemma 1

To prove Lemma 1, we provide a more refined asymptotic analysis as follows

$$\bar{R}_{\text{sym}} = \begin{cases} \Theta(1/K), & \text{if } \eta \in [-1, 0], \\ \Theta(K^{\eta-1}), & \text{if } \eta \in (0, 1], \\ (\eta - 1)\log(K) + O(1), & \text{if } \eta \in (1, \infty), \end{cases} \tag{30}$$

with high probability (*w.h.p.*).

1) Non-increasing per-user power regime $\eta \in [-1, 0]$*:* In this regime, $P = \Theta(K^{\eta+1})$, and $p = \Theta(K^{\eta})$ is vanishing or fixed. The CSIT error is $\sigma^2 = \Theta(1)$. In consistence with (30), we shall show the detailed scaling

$$\begin{cases} \bar{R}_{\text{sym}} = \Theta(\frac{1}{K}), & \text{if } \eta = -1, \\ (\eta+1)\Theta(\frac{\log(K)}{K}) \geq \bar{R}_{\text{sym}} \geq \Theta(\frac{1}{K}), & \text{if } \eta \in (-1, 0]. \end{cases} \tag{31}$$

From the fact that $\frac{x}{1+x} \leq \ln(1+x) \leq x$ for $x > -1$, we have

$$\mathbb{E}\left[\frac{\text{SINR}_k}{1+\text{SINR}_k}\right] \leq \frac{\bar{R}_{\text{sym}}}{\ln(2)} \leq \mathbb{E}\left[\text{SINR}_k\right]. \tag{32}$$

First, $\frac{\text{SINR}_k}{1+\text{SINR}_k} = \frac{|\text{G}_k|^2}{p^{-1}+|\text{G}_k|^2+\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2}$ and from the Jensen's inequality for the log function

$$\log \mathbb{E}\left[\frac{\text{SINR}_k}{1+\text{SINR}_k}\right] \geq \mathbb{E}\left[\log\left(\frac{\text{SINR}_k}{1+\text{SINR}_k}\right)\right] \tag{33}$$

$$= \mathbb{E}\left[\log\left(|\text{G}_k|^2\right)\right] - \mathbb{E}\left[\log\left(\frac{1}{p} + |\text{G}_k|^2 + \sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2\right)\right] \tag{34}$$

$$\geq \mathbb{E}[\log\left(|\text{G}_k|^2\right)] - \log\left(\frac{1}{p} + \mathbb{E}\left[|\text{G}_k|^2\right] + \mathbb{E}\left[\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2\right]\right). \tag{35}$$

Since $|\text{G}_k|^2$ is exponentially distributed with mean 1, $\ln(|\text{G}_k|^2)$ follows the generalized extreme value distribution of location 0, scale -1 and shape 0, denoted as GEV$(0, -1, 0)$. Thus $\mathbb{E}\left[\log(|\text{G}_k|^2)\right] = -\frac{\psi}{\ln(2)}$ where $\psi \simeq 0.577$ is the Euler's constant. Moreover, $\mathbb{E}\left[\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2\right] = (K-1)\sigma^2$. Then

$$\mathbb{E}\left[\frac{\text{SINR}_k}{1+\text{SINR}_k}\right] \geq \frac{2^{-\psi/\ln(2)}}{p^{-1}+1+(K-1)\sigma^2} = \Theta(1/K) \tag{36}$$

since $p^{-1} = \Theta(K^{-\eta})$ is dominated by $1+(K-1)\sigma^2 = \Theta(K)$.

Next, we focus on the upper bound $\mathbb{E}\left[\text{SINR}_k\right]$. If $\eta = -1$, it is simply

$$\mathbb{E}\left[\text{SINR}_k\right] \leq \mathbb{E}\left[|\text{G}_k|^2 p\right] = p\mathbb{E}\left[|\text{G}_k|^2\right] = \Theta(1/K). \tag{37}$$

If $\eta \in (-1, 0]$, for any δ, we can write that

$$\mathbb{E}\left[\text{SINR}_k\right] \leq \mathbb{E}\left(\frac{1}{p^{-1}+\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2}\right) \quad \textit{w.h.p.} \tag{38}$$

by considering the event $|\text{G}_k|^2 < K^{\delta}$ occurring *w.h.p.*, and choosing δ close to 0. Denote $\mathbf{W}_{-k} := [\mathbf{W}_1 \dots \mathbf{W}_{k-1}\mathbf{W}_{k+1} \dots \mathbf{W}_K]^{\text{T}} \in \mathbb{C}^{(K-1)\times n_{\text{t}}}$ $(n_t \geq K)$, then

$$\frac{1}{\sigma^2}\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2 = \frac{1}{\sigma^2}\|\tilde{\mathbf{H}}_k\mathbf{W}_{-k}\|^2 = \frac{1}{\sigma^2}\tilde{\mathbf{H}}_k^{\text{H}}\mathbf{W}_{-k}^{\text{H}}\mathbf{W}_{-k}\tilde{\mathbf{H}}_k$$

$$= \frac{1}{\sigma^2}\tilde{\mathbf{H}}_k^{\text{H}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\text{H}}\tilde{\mathbf{H}}_k = \mathbf{J}_k^{\text{H}}\mathbf{\Lambda}\mathbf{J}_k = \sum_{i=1}^{K-1}\lambda_i|J_{k,i}|^2$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1 \dots \lambda_{K-1})$ with $\sum_i \lambda_i = K-1$ and $\mathbf{J}_k$ is identically distributed to $\tilde{\mathbf{H}}_k/\sigma$. Given $\{\lambda_i\}$, $\frac{1}{\sigma^2}\sum_{l\neq k}|\tilde{\text{G}}_{k,l}|^2$ is the sum of $K-1$ independent exponential random variables, each of mean λ_i, $i = 1, \dots, K-1$. From (38), we can derive

$$\mathbb{E}\left[\text{SINR}_k\right] \leq \mathbb{E}\left[\frac{1}{p^{-1}+\sigma^2\sum_{i=1}^{K-1}\lambda_i|J_{k,i}|^2}\right] \tag{39}$$

$$= \frac{1}{K-1}\mathbb{E}\left[\frac{1}{\sigma^2\frac{1}{K-1}\sum_{i=1}^{K-1}\left(\frac{p^{-1}}{(K-1)\sigma^2}+\lambda_i|J_{k,i}|^2\right)}\right] \tag{40}$$

$$\leq \frac{1}{K-1}\mathbb{E}\left[\frac{1}{\sigma^2\prod_{i=1}^{K-1}\left(\frac{p^{-1}}{(K-1)\sigma^2}+\lambda_i|J_{k,i}|^2\right)^{\frac{1}{K-1}}}\right] \tag{41}$$

$$= \frac{1}{(K-1)\sigma^2}\mathbb{E}\left[\prod_{i=1}^{K-1}\left(\frac{1}{\frac{p^{-1}}{(K-1)\sigma^2}+\lambda_i|J_{k,i}|^2}\right)^{\frac{1}{K-1}}\right] \tag{42}$$

$$\leq \frac{1}{(K-1)\sigma^2}\frac{1}{K-1}\mathbb{E}\left[\sum_{i=1}^{K-1}\frac{1}{\frac{p^{-1}}{(K-1)\sigma^2}+\lambda_i|J_{k,i}|^2}\right] \tag{43}$$

$$= \frac{1}{(K-1)\sigma^2}\mathbb{E}\left[\frac{1}{\frac{p^{-1}}{(K-1)\sigma^2}+\lambda|J_k|^2}\right] \tag{44}$$

where the expectation is over both $\{\lambda_i\}$ and $\{J_{k,i}\}$. In (41) and (43) we apply the AM-GM inequality and in the last line, J_k is exponentially distributed with mean 1. Then

$$\mathbb{E}\left[\text{SINR}_k \mid \lambda\right] \leq \frac{1}{(K-1)\sigma^2}\mathbb{E}_{J_k}\left(\frac{1}{\frac{p^{-1}}{(K-1)\sigma^2}+\lambda|J_k|^2}\right) \tag{45}$$

$$= \frac{1}{(K-1)\sigma^2}\frac{\exp\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)}{\lambda}\text{E}_1\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right) \tag{46}$$

where $\text{E}_1(x) := \int_x^{\infty}\frac{e^{-t}}{t}dt$ is the exponential integral function with the property $\frac{1}{2}e^{-x}\ln\left(1+\frac{2}{x}\right) \leq \text{E}_1(x) \leq e^{-x}\ln\left(1+\frac{1}{x}\right)$. Hence in large K regime, $\text{E}_1\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)$ scales as $\Theta\left(\log\left(\frac{(K-1)\sigma^2\lambda}{p^{-1}}\right)\right) = (\eta+1)\Theta(\log(K))$, and $\exp\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)$ goes to 1. Thus, we have

$$\mathbb{E}\left[\text{SINR}_k \mid \lambda\right] \leq (\eta+1)\Theta(\log(K)/K) \tag{47}$$

for all realization of λ. This holds for $\mathbb{E}\left[\text{SINR}_k\right]$ in general.

Substituting the bounds (37) and (47) of $\mathbb{E}\left[\text{SINR}_k\right]$ and (36) of $\mathbb{E}\left[\frac{\text{SINR}_k}{1+\text{SINR}_k}\right]$ to (32), (31) is proved.

2) Increasing per-user power regime $\eta > 0$*:* In this regime, $p = \Theta(K^{\eta})$ for some $\eta > 0$, $\sigma^2 = \Theta(p^{-1}) = \Theta(K^{-\eta})$. We shall show the following scaling, in consistence with (30)

$$\begin{cases} \Theta(K^{\eta-1+\phi}) \geq \bar{R}_{\text{sym}} \geq \Theta(K^{\eta-1}), & \text{if } 1 > \eta > 0 \\ \log\log(K) + O(1) \geq \bar{R}_{\text{sym}} \geq \Theta(1), & \text{if } \eta = 1 \\ (\eta-1+\phi)\log(K) + O(1) \geq \bar{R}_{\text{sym}} \geq (\eta-1)\log(K) & \\ \qquad\qquad +O(1), & \text{if } \eta > 1 \end{cases} \tag{48}$$

for any $\phi > 0$.

First we derive the bounds for $\mathbb{E}[\mathrm{SINR}_k]$. By the same argument as with $\mathbb{E}\left[\frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}\right]$, we have

$$\mathbb{E}[\mathrm{SINR}_k] \geq \frac{2^{-\psi/\ln(2)}}{p^{-1}+(K-1)\sigma^2} \geq \Theta(K^{\eta-1}) \tag{49}$$

since both p^{-1} and σ^2 scale as $\Theta(K^{-\eta})$. In addition, the following is still valid in this regime

$$\mathbb{E}[\mathrm{SINR}_k \mid \lambda] \leq \frac{1}{(K-1)\sigma^2\lambda}\exp\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)\mathrm{E}_1\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right).$$

In the righthand side, when K is large, $\frac{1}{(K-1)\sigma^2\lambda}$ scales as $\Theta(K^{\eta-1})$, $\exp\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)$ goes to 1, and $\mathrm{E}_1\left(\frac{p^{-1}}{(K-1)\sigma^2\lambda}\right)$ scales as $\Theta(\log(K))$. Consequently, $\mathbb{E}[\mathrm{SINR}_k \mid \lambda] \leq \Theta(K^{\eta-1}\log(K)) \leq \Theta(K^{\eta-1+\phi})$ for any $\phi > 0$. This holds as an upper bound for $\mathbb{E}[\mathrm{SINR}_k]$ also.

Hence, $\Theta(K^{\eta-1+\phi}) \geq \mathbb{E}[\mathrm{SINR}_k] \geq \Theta(K^{\eta-1})$ for any $\phi > 0$. For $\bar{R}_{\mathrm{sym}}$, we look at three following cases.

- $0 < \eta < 1$: $\mathbb{E}[\mathrm{SINR}_k]$ vanishes as K is large. Again, we have the bounds
$$\mathbb{E}\left[\frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}\right] \leq \frac{\bar{R}_{\mathrm{sym}}}{\ln(2)} \leq \mathbb{E}[\mathrm{SINR}_k]. \tag{50}$$
where
$$\mathbb{E}\left[\frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}\right] \geq \frac{2^{-\psi/\ln(2)}}{p^{-1}+1+(K-1)\sigma^2} = \Theta(K^{\eta-1}). \tag{51}$$
Hence $\Theta(K^{\eta-1+\phi}) \geq \bar{R}_{\mathrm{sym}} \geq \Theta(K^{\eta-1})$.
- $\eta = 1$: $\Theta(\log(K)) \geq \mathbb{E}[\mathrm{SINR}_k] \geq \Theta(1)$, $\mathbb{E}[\mathrm{SINR}_k]$ scales up with K and $\log\log(K)+O(1) \geq \bar{R}_{\mathrm{sym}} \geq \Theta(1)$.
- $\eta > 1$: $\mathbb{E}[\mathrm{SINR}_k]$ scales up and $(\eta-1+\phi)\log(K) + O(1) \geq \bar{R}_{\mathrm{sym}} \geq (\eta-1)\log(K)+O(1)$.

Hence (48) is proved. It and (31) constitute the refined scaling (30) of $\bar{R}_{\mathrm{sym}}$.

B. Proof of Lemma 2

When $n_t = 1$, $\min_k\{\mathrm{SNR}_k^{(0)}\} = P\min_k\{\|\mathbf{H}_k\|^2\}$. As we assume i.i.d. Rayleigh fading channel, $\min_k\{\|\mathbf{H}_k\|^2\}$ is the minimum of K independent exponentially distributed random variables each with mean 1. Then $\min_k\{\|\mathbf{H}_k\|^2\}$ is also exponentially distributed with parameter $\sum_1^K 1 = K$. Hence $\min_k\{\mathrm{SNR}_k^{(0)}\}$ exponentially distributed with mean P/K.

Now we consider $n_t \geq (1+\delta)\log(K)$ for any $\delta \geq 0$. First, we compute the upper bound for $\min_k \frac{\|\mathbf{H}_k\|^2}{n_t}$. Applying the Chernoff bound to the random variable $\frac{\|\mathbf{H}_k\|^2}{n_t}$, we can show that for any $\nu \geq 0$ and any ϵ

$$\Pr\left(\frac{\|\mathbf{H}_k\|^2}{n_t} \geq 1+\epsilon\right) \leq e^{-\nu(1+\epsilon)}\left(1-\frac{\nu}{n_t}\right)^{-n_t}. \tag{52}$$

Choosing the value of ν which minimizes the upper bound, which is $\nu = n_t\frac{\epsilon}{1+\epsilon}$ for $\epsilon \geq 0$, yeilds

$$\Pr\left(\frac{\|\mathbf{H}_k\|^2}{n_t} \geq 1+\epsilon\right) \leq e^{n_t(-\epsilon+\log(1+\epsilon))}. \tag{53}$$

This is used to bound the probability $\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \geq 1+\epsilon\right)$

$$\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \geq 1+\epsilon\right) \leq e^{Kn_t(-\epsilon+\log(1+\epsilon))} \tag{54}$$

This probability vanishes for any $\epsilon \geq 0$ such that $-\epsilon+\log(1+\epsilon) < 0$, or equivalently $\epsilon > 0$. Thus

$$\lim_{K\to\infty}\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1+\epsilon\right) = 1 \text{ for any } \epsilon > 0. \tag{55}$$

Now we compute the lower bound for $\min_k \frac{\|\mathbf{H}_k\|^2}{n_t}$. We apply the Chernoff bound again to yeild

$$\Pr\left(\frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1-\epsilon\right) \leq e^{\nu(1+\epsilon)}\left(1+\frac{\nu}{n_t}\right)^{-n_t} \tag{56}$$

for any $\nu \geq 0$ and any ϵ. Then with the optimal $\nu = n_t\frac{\epsilon}{1-\epsilon}$ for $\epsilon \leq 1$, which minimizes the above upper bound, we have

$$\Pr\left(\frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1-\epsilon\right) \leq e^{n_t(\epsilon+\log(1-\epsilon))}. \tag{57}$$

Hence we obtain the bound for $\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1-\epsilon\right)$

$$\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1-\epsilon\right) \leq 1-(1-e^{n_t(\epsilon+\log(1-\epsilon))})^K \tag{58}$$

$$< Ke^{n_t(\epsilon+\log(1-\epsilon))}. \tag{59}$$

where for the last line we use the fact that $(1-x)^K > 1-Kx$ for any $x \in (0,1)$ and consider the set $\{\epsilon : \epsilon+\log(1-\epsilon) < -1\}$. The infimum of this set is $\epsilon_0 \simeq 0.8414$. Then

$$\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \leq 1-\epsilon_0\right) < Ke^{-n_t} < Ke^{-(1+\delta)\log(K)} = K^{-\delta}. \tag{60}$$

This probability vanishes as K grows since $\delta \geq 0$ by assumption. Hence we can write that

$$\lim_{K\to\infty}\Pr\left(\min_k \frac{\|\mathbf{H}_k\|^2}{n_t} \geq 1-\epsilon_0\right) = 1. \tag{61}$$

The (55) and (61) constitute (16) and conclude the proof.

C. Proof of Lemma 3

A more refined asymptotic analysis of $\bar{R}_0$ is as follows

$$\bar{R}_0 = \begin{cases}\eta\log(K)+O(1), & \text{if } n_{\mathrm{t}} = 1,\\ (\eta+1)\log(K)+O(1), & \text{if } n_{\mathrm{t}} = \Omega(\log(K)).\end{cases} \tag{62}$$

First, when $n_{\mathrm{t}} = 1$, $\min_k\{\mathrm{SNR}_k^{(0)}\}$ is exponentially distributed with mean P/K from Lemma 2. Then

$$\bar{R}_0 = \mathbb{E}\left[\log\left(\min_k\{\mathrm{SNR}_k^{(0)}\}\right)\right] + O(1) \tag{63}$$

$$= \frac{\mathbb{E}\left[\ln\left(\min_k\{\mathrm{SNR}_k^{(0)}\}\right)\right]}{\ln(2)} + O(1) \tag{64}$$

$$= \frac{\ln(P/K)-\psi}{\ln(2)} + O(1) \tag{65}$$

$$= \eta\log(K)+O(1) \tag{66}$$

where we use the fact that $\ln\left(\frac{\min_k\left\{\mathrm{SNR}_k^{(0)}\right\}}{P/K}\right) = \ln\left(\min_k\left\{\mathrm{SNR}_k^{(0)}\right\}\right) - \ln(P/K)$ follows the GEV$(0,-1,0)$ distribution with mean $-\psi$, with ψ the Euler's constant.

Next we consider the case $n_\mathrm{t} = \Omega(\log(K))$. It follows from (16) in Lemma 2 that

$$\bar{R}_0 \in [\log(1+P(1-\epsilon_0)), \log(1+P(1+\epsilon))] \quad \textit{w.h.p.} \quad (67)$$

where $\epsilon_0 \simeq 0.8414$ and $\epsilon > 0$ can be arbitrarily small. Thus $\bar{R}_0 = \log(P) + O(1) = (\eta+1)\log(K) + O(1)$.

D. Asymptotic behaviors of $R_{\textit{uni-c}}$ and $R_{\textit{mul-c}}$ in large K regime

From the scaling of $\bar{R}_{\mathrm{sym}}$ and $\bar{R}_0$ given in Lemma 1 and Lemma 3, respectively, we can have the symptotic behaviors of $R_{\text{uni-c}}$ and $R_{\text{mul-c}}$ when K is large in some regimes of the total power P as follows.

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

$$R_{\text{uni-c}} = \Theta\big(\frac{1}{1-m}\big), \quad (68)$$

$$R_{\text{mul-c}} = \Theta\left(\frac{1+Km}{1-m}\right). \quad (69)$$

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

$$R_{\text{uni-c}} = \Theta\big(\frac{1}{1-m}\big), \quad (70)$$

$$R_{\text{mul-c}} = \frac{1+Km}{1-m}\log(K) + O(1). \quad (71)$$

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

$$R_{\text{uni-c}} = \begin{cases} \Theta\big(\frac{1}{1-m}K^\eta\big), & \text{if } \eta \le 1 \\ \frac{1}{1-m}(\eta-1)K\log(K) + O(1), & \text{if } \eta > 1 \end{cases} \quad (72)$$

$$R_{\text{mul-c}} = \frac{1+Km}{1-m}(\eta+1)\log(K) + O(1). \quad (73)$$

These scalings are straightforward since $R_{\text{uni-c}}$ and $R_{\text{mul-c}}$ are linear functions of $\bar{R}_{\mathrm{sym}}$ and $\bar{R}_0$, respectively.

E. Proof of Proposition 2

Recall that $P = \Theta(K^{\eta+1})$. Let $P - P_0 = \Theta(K^{\beta+1})$ for some $\beta \le \eta$. Then $P_0 = \Theta(K^{\eta+1})$ and $\frac{P-P_0}{K} = \Theta(K^\beta), k = 1, \ldots, K$. Then the Proposition 2 can be interpreted as follows. The value of β which achieve the optimal scaling of R_{mix} is

$$\beta = \begin{cases} -1, & \text{if } \eta \in \left[-1, \frac{1}{1-m}\right], \\ \eta, & \text{if } \eta \in \left(\frac{1}{1-m}, \infty\right). \end{cases} \quad (74)$$

First, we notice that for a given η, we have no interest of letting the private power decrease with K, i.e. $\beta < -1$. Thus, we look into the following regimes and can derive the scaling of $R_{\text{mul-c}}^{\text{mix}}$ and $R_{\text{uni-c}}^{\text{mix}}$ by using the Lemma 2 and establishing the upper and lower bounds as in the proof of Lemma 1 and Lemma 3. Details are omitted here for brevity.

1) Non-increasing per-user total power regime $\eta \in [-1, 0]$: In this regime $-1 \le \beta \le \eta \le 0$, $\sigma^2 = \Theta(1)$. Then as $K \to \infty$

$$R_{\text{uni-c}}^{\text{mix}} = \Theta\left(\frac{1}{1-m}\right), \quad (75)$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}\log(1+K^{\eta-\beta}) + O(1). \quad (76)$$

We see that in this regime, the private rate is negligible to the common rate and has no contribution to the scaling of R_{mix}. In addition, the scaling of $R_{\text{mul-c}}^{\text{mix}}$ is decreasing with β. Hence the optimal value of β is -1.

2) Increasing per-user total power regime $\eta > 0$: In this regime, the total power per user P/K scales up with K. We consider two sub-regimes in term of per-user private power, and derive the local optimal rate in each sub-regime as follows.

a) Non-increasing per-user private power regime $-1 \le \beta \le 0$: In this sub-regime, the CSIT error is still $\sigma^2 = \Theta(1)$. The scaling of $R_{\text{mul-c}}^{\text{mix}}$ and $R_{\text{uni-c}}^{\text{mix}}$ is

$$R_{\text{uni-c}}^{\text{mix}} = \Theta\left(\frac{1}{1-m}\right), \quad (77)$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}(\eta-\beta)\log(K) + O(1). \quad (78)$$

Again, we see that the private rate $R_{\text{uni-c}}^{\text{mix}}$ is dominated and does not contribute to the scaling of the total sum rate. The scaling of $R_{\text{mul-c}}^{\text{mix}}$ still decreases with β. Hence the local optimal value of β in this case is -1 and the local optimal sum rate is

$$R_{\text{mix}} = \frac{1+Km}{1-m}(\eta+1)\log(K) + O(1). \quad (79)$$

b) Increasing per-user private power regime $\eta \ge \beta > 0$: In this sub-regime, the estimation error decreases with $\frac{P-P_0}{K}$ as $\sigma^2 = \Theta(\big(\frac{P-P_0}{K}\big)^{-1}) = \Theta(K^{-\beta})$. The scaling is

$$R_{\text{uni-c}}^{\text{mix}} = \begin{cases} \Theta\left(\frac{1}{1-m}K^\beta\right), & \text{if } 0 < \beta \le 1 \\ \frac{1}{1-m}(\beta-1)K\log(K) + O(1), & \text{if } \beta > 1 \end{cases} \quad (80)$$

$$R_{\text{mul-c}}^{\text{mix}} = \frac{1+Km}{1-m}(\eta - (\beta-1)^+)\log(K) + O(1). \quad (81)$$

We can see that, provided that m is nonzero and fixed, the private rate is still dominated by the common rate if $0 < \beta \le 1$. The scaling of $R_{\text{mul-c}}^{\text{mix}}$ is independent of β given that $0 < \beta \le 1$ and increasing with β given that $1 < \beta \le \eta$. Hence, the local optimal β in this case is η and the local optimal sum rate is

$$R_{\text{mix}} = \frac{1}{1-m}\left[(\eta-1)K + (1+Km)\right]\log(K) + O(1). \quad (82)$$

Comparing two local optimal rates (79) and (82), we have the global optimal β in increasing per-user total power regime is -1 if $0 < \eta \le \frac{1}{1-m}$ and η if $\eta > \frac{1}{1-m}$.

Finally, by summarizing the optimal β in the two regimes above, we conclude the proof.