# Optimal Power Allocation in Cache-Aided Non-Orthogonal Multiple Access Systems

Khai Nguyen Doan*, Wonjae Shin‡, Mojtaba Vaezi†, H. Vincent Poor† and Tony Q. S. Quek*
*Information Systems Technology and Design, Singapore University of Technology and Design, Singapore
Email: nguyenkhai_doan@mymail.sutd.edu.sg, tonyquek@sutd.edu.sg
†Department of Electrical Engineering, Princeton University, Princeton, NJ, USA
Email: {mvaezi, poor}@princeton.edu
‡Department of Electronics Engineering, Pusan National University, Pusan, South Korea
Email: wjshin@pnu.ac.kr

***Abstract*—This work combines non-orthogonal multiple access (NOMA) and caching, two prominent techniques for future communication networks. Specifically, a cache-aided cellular network with Rayleigh fading channels is analyzed. By exploiting the channel distribution, users requests, and cache contents in given time, an optimal power allocation policy for the superposed signal is derived. The goal is to maximize the system success probability, i.e., the probability that all users can successfully decode their desired signals. The analysis highlights the benefits of introducing caching to NOMA-based systems. Simulation results confirm the analysis and demonstrate the efficiency of the proposed power allocation.**

## I. Introduction

Future communication networks must satisfy a dramatic increase in data demand due to the growth of electronic devices [1]. The system capacity and user quality of service (QoS) can be improved by deploying networks with a higher density of access points. This, in turn, causes massive load on the backhaul [2]. Caching, i.e., pushing contents to user's devices [3]–[6], appears to be a promising solution for this challenge. This technique offers the users a chance to retrieve the desired contents from their own devices, thus, helping avoid unnecessary transmissions.

Non-orthogonal multiple access (NOMA) is another potential candidate to improve the system capacity and user experience. This technique has shown to be more efficient than its counterpart, orthogonal multiple access (OMA), in term of reducing the spectrum usage and enhancing the power efficiency [7]–[10]. In the power domain, NOMA supports the communication of multiple users using the same radio resource in time and frequency by superimposing the users signal [11], [12]. Then, the successive interference cancellation (SIC) can be applied at the receiver for signal decoding.

Currently, there are very few works investigating potential benefits of caching in the context of NOMA systems. One example is [13] which considers a network consisting of a base station (BS) and several content servers where the system operation relies on a combination of caching and NOMA. Caches are implemented at the content servers and NOMA is applied whenever data needs to be delivered from the BS to content servers or from content servers to mobile users. However, the power allocation policy is to guarantee that the most popular file can be successfully delivered from the BS to a predefined number of content servers instead of being optimized to reduce the delivery outage probability at the user side. In addition, the use of cache contents to eliminate interference is not considered when applying SIC. From our point of view, finding an optimal power allocation to suppress the outage probability or equivalently, maximize the success probability, is important. Since it plays a key role in improving the user's QoS, while ensuring the fairness. Specifically, the success probability is defined as the probability that all users can successfully decode their desired signals. In addition, with the involvement of caching, this problem has not been fully addressed. For example, in NOMA without caching, the power is conventionally allocated in an inverse order of channel gains between users and the BS. Nevertheless, with the aid of caching, users can remove signals other than their desired ones from the superposed signal using the cache contents. Therefore, the conventional power allocation scheme may not be optimal anymore, and this inspires finding a more appropriate method.

In this paper, we derive an optimal power allocation policy for the cache-aided NOMA system to maximize the success probability. Importantly, the success probability, by its definition, can ensure the fairness among users. Besides, our design exploits the knowledge of channel gain distribution, the requests and cache contents of users in an instant time. This is done with the assumption that, users are paired and in the worst case, we can afford to allocate orthogonal subchannels to user pairs. In addition, numerical results are given to gain more insight to the system operation as well as the relation between caching and NOMA.

The remainder of this work is organized as follows. Section II describes the system model under consideration. Section III presents the optimal power allocation. Finally, Section IV describes our experimental setup and simulation results to validate our theoretical results.

## II. System Model

We consider a system consisting of $K$ users served by a BS as in Fig. 1. Each user's device has a cache with finite capacity and we assume that users cache a file as a whole without partitioning. As will be seen in the next section, the total number of files that a user can cache does not directly

978-1-5386-4328-0/18/$31.00 ©2018 IEEE

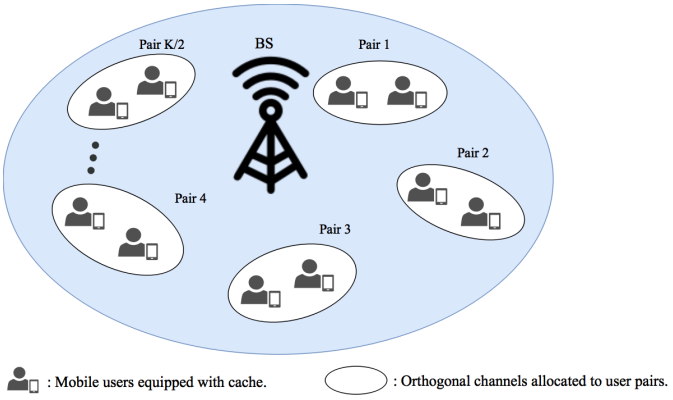


Fig. 1. A BS serving $K$ mobile users who are equipped with caches. Users are paired and allocated orthogonal subchannels.

involve in the power allocation decision. Thus, we just assume that this quantity is finite instead of specifying it for simplicity. Typically, with caching, users fetch and store a set of files during the off-peak time called *caching phase*. In this work, we assume that the caching phase has already taken place to consider the next stage called *requesting phase*. In this phase, each user is assumed to request for a file in the library. In addition, we assume that the BS has information about files stored on each device. For example, when sending the request messages to BS, users can add information saying which files currently present in their caches.

With NOMA, when a superposed signal is transmitted from the BS to users, the SIC scheme is used at each mobile user to decode the desired signal. As discussed in [10] and the references therein, SIC may increase the complexity at the receivers as there are more users occupying the same subchannel. Therefore, it is reasonable to restrict the number of users associated with each subchannel to two. In this work, we assume that users are paired and orthogonal subchannels are allocated to user pairs.[1]

Considering a single subchannel occupied by two users, let us denote $h_1$ (and $h_2$) to be the channel coefficient between user 1 (and user 2) and the BS where $|h_1|$ (and $|h_2|$) follows Rayleigh distribution with parameter $\sigma_1^2$ (and $\sigma_2^2$). Then, $|h_1|^2$ and $|h_2|^2$ will follow an exponential distribution with parameter $\lambda_1 = 2/\sigma_1^2$ and $\lambda_2 = 2/\sigma_2^2$, respectively. We denote $d_1$ and $d_2$ to be the distance from user 1 and 2 to the BS, respectively. Then, the signal received at user 1 and 2 will be

$$y_1 = \frac{h_1}{d_1^\gamma}\left(x_1\sqrt{\alpha} + x_2\sqrt{1-\alpha}\right) + n_1 \tag{1}$$

and

$$y_2 = \frac{h_2}{d_2^\gamma}\left(x_1\sqrt{\alpha} + x_2\sqrt{1-\alpha}\right) + n_2 \tag{2}$$

[1]For this model, the subchannels can be allocated dynamically by grouping users having the same requested file and allocate them a single subchannel to trigger the broadcasting feature of wireless networks. The situation of two users per subchannel as our consideration is actually the worst case when there are no more than two users requesting the same content.

respectively. $x_1$ and $x_2$ are the signals corresponding to requested files of user 1 and 2, respectively, $\alpha$ is the portion of the transmission power $P$ allocated to the signal of the user 1, $\gamma$ is the path loss coefficient, and $n_1$ is Gaussian noise with mean 0 and variance $\sigma^2$. Denote $\epsilon_1$ and $\epsilon_2$ as the minimum SINR for which the requested files of user 1 and 2, respectively, can be decoded. Let the SNR be $\rho = \frac{P}{\sigma^2}$. Without loss of generality, we assume that $\frac{\zeta_1}{\zeta_2} > 1$ where $\zeta_1 = \lambda_1\epsilon_1\beta_1$, $\zeta_2 = \lambda_2\epsilon_2\beta_2$, $\beta_1 = \frac{d_1^\gamma}{\rho}$ and $\beta_2 = \frac{d_2^\gamma}{\rho}$. Note that the larger $\zeta_1$ (or $\zeta_2$) is, the harder for user 1 (or user 2) to successfully decode its desired signal. Therefore, this assumption implies that, without caching, user 1 has a worse condition to be satisfied.

In SIC without caching, many users may need to decode a sequence of signals before obtaining their desired ones. In this case, the failure can occur if users fail to decode one of those signals. However, with caching, users can remove or reduce the interference without decoding using cache contents, which increases the success probability. In this case, the optimal power allocation policy will not only depend on the channel information but also the requests and cache contents of users in an instant time. Therefore, we aim to design a power allocation policy that can take into account all of this information to maximize the success probability.

## III. Optimal Power Allocation

In our work, the power allocation consists of two stages. The first stage is to share the total transmission power to user pairs. Then, the second stage is to allocate portions of a given power amount to users in each pair. However, due to the dependence between these stages, we will first present the second stage in subsection III-A, then, the first stage in subsection III-B. Note that this two-stage process is optimized as each stage is optimized separately. This is because user pairs are isolated in term of interference due to orthogonal subchannels allocation.

### A. A user pair with a single subchannel

In this subsection, we deal with a user pair occupying a single subchannel. Note that the SIC decoding order depends on which signal is given more power, i.e., $0.5 \leq \alpha \leq 1$ or $0 \leq \alpha \leq 0.5$, and so does the objective function. Therefore, we will need to optimize two probability functions in both ranges of $\alpha$. As mentioned in the previous section, a limited number of situations can arise. In some of these cases, the optimal power allocation is trivial whereas in some other cases, the problem is involved. We first consider the trivial cases before listing the others.

1) When both users can find their requested files in their caches, thus, no transmission is required.
2) When only one of the user's demand can be found in the cache, hence to maximize the success probability, all of power $P$ will be used for the transmission to the unsatisfied user, i.e., $\alpha = 0$ assuming user 2 is the one has not been satisfied yet. Then the desired file is successfully decoded when

$$|h_2|^2 / \beta_2 \geq \epsilon_2 \tag{3}$$

or equivalently,

$$|h_2|^2 \geq \epsilon_2 \beta_2 \tag{4}$$

Since $|h_2|^2$ has exponential distribution, the success probability will be

$$\Pr\left(|h_2|^2 \geq \epsilon_2 \beta_2\right) = \exp\left(-\lambda_2 \epsilon_2 \beta_2\right) \tag{5}$$

3) When users request for the same file, e.g. file $f$, but neither of them have cached it. Then, a single signal corresponding to that file is broadcasted with power $P$. Similar to the above, the desired file is successfully decoded at each user when

$$|h_1|^2 / \beta_1 \geq \epsilon_f \tag{6}$$

$$|h_2|^2 / \beta_2 \geq \epsilon_f \tag{7}$$

where $\epsilon_f$ is the minimum SINR to successfully decode the file $f$, and the success probability is given by

$$\begin{aligned} &\Pr\left(|h_1|^2 \geq \epsilon_f \beta_1, |h_2|^2 \geq \epsilon_f \beta_2\right) \\ &= \exp\left(-\lambda_1 \epsilon_f \beta_1 - \lambda_2 \epsilon_f \beta_2\right) \end{aligned} \tag{8}$$

Without loss of generality, we assume that user 1 request for file $f_1$ and user 2 requests for $f_2$. Then, four cases which are more complicated are listed as follows

- *Case 1:* User 1 has cached $f_2$, user 2 has had a cache miss.
- *Case 2:* User 1 has had a cache miss, user 2 has cached $f_1$.
- *Case 3:* User 1 has cached $f_2$, user 2 has cached $f_1$.
- *Case 4:* Both users have had cache misses.

where a user experiences a cache miss when it does not cache its own desired file and that of the other user. We will analyze each case separately and derive the optimal power allocation. Starting with the first case, since user 1 has already cached the requested file of user 2, it can remove the interference without decoding. Hence, user 1 can decode its desired content when

$$|h_1|^2 \alpha / \beta_1 \geq \epsilon_1 \tag{9}$$

Then, we have the following conditions for successful decoding at user 2 for $\alpha \geq 0.5$ and $\alpha < 0.5$, respectively,

$$|h_2|^2 \alpha \Big/ \left(|h_2|^2 (1-\alpha) + \beta_2\right) \geq \epsilon_1 \tag{10a}$$

$$|h_2|^2 (1-\alpha) / \beta_2 \geq \epsilon_2 \tag{10b}$$

and

$$|h_2|^2 (1-\alpha) \Big/ \left(|h_2|^2 \alpha + \beta_2\right) \geq \epsilon_2 \tag{11}$$

The success probability in the former and later cases are expressed by

$$f_1^{Case1}(\alpha) = \begin{cases} \exp\left(-\frac{\zeta_1}{\alpha} - \max\left(\frac{\lambda_2 \epsilon_1 \beta_2}{(1+\epsilon_1)\alpha - \epsilon_1}, \frac{\zeta_2}{1-\alpha}\right)\right) \\ \qquad\qquad , \alpha > 1 - \frac{1}{1+\epsilon_1} \\ 0 \qquad\qquad , \text{otherwise} \end{cases} \tag{12}$$

and

$$f_2^{Case1}(\alpha) = \begin{cases} \exp\left(-\frac{\zeta_1}{\alpha} - \frac{\zeta_2}{1-(1+\epsilon_2)\alpha}\right) & , \alpha < \frac{1}{1+\epsilon_2} \\ 0 & , \text{otherwise} \end{cases} \tag{13}$$

respectively. To this end, the optimal power allocation policy is given in Theorem 1.

*Theorem 1:* For case 1, the success probability is maximized when

$$\alpha = \begin{cases} x_1^{Case1} & , \text{if } f_1^{Case1}\left(x_1^{Case1}\right) \leq f_2^{Case1}\left(x_2^{Case1}\right) \\ x_2^{Case1} & , \text{otherwise} \end{cases} \tag{14}$$

where $x_1^{Case1}$ and $x_2^{Case1}$ depend on the ratio $\zeta = \frac{\zeta_1}{\zeta_2} > 1$ in the following way

$$x_1^{Case1} = \max\left(1 - \frac{1}{\sqrt{\zeta}+1}, 1 - \frac{1}{1+\epsilon_1+\frac{\epsilon_1}{\epsilon_2}}\right) \tag{15}$$

$$x_2^{Case1} = \min\left(\frac{1}{1+\epsilon_2}\left(1 - \frac{1}{\sqrt{\zeta(1+\epsilon_2)}+1}\right), 0.5\right) \tag{16}$$

The above results are obtained by maximizing the success probability in two cases when $0.5 \leq \alpha \leq 1$ and $0 \leq \alpha \leq 0.5$ separately. For the former case, we minimize the argument of the exponential function in (12) whose solution is the first argument of the $\max$ function in (15). Then, by combining it with the constraints $\alpha > 1 - \frac{1}{1+\epsilon_1}$ and $0.5 \leq \alpha \leq 1$, we obtain (15). Similarly for the case $0 \leq \alpha \leq 0.5$ whose solution is (16). To this end, we can plug (15) and (16) into the corresponding objective function for comparison and choose the better one, which leads to the result in Theorem 1. The full proof of this as well as the subsequent ones are omitted due to the space limitation. However, they can be found in a longer version of this work.

Next, for case 2, the condition for user 2 to successfully decode its own signal is

$$|h_2|^2 (1-\alpha) / \beta_2 \geq \epsilon_2 \tag{17}$$

and the remaining conditions for the success event when $\alpha$ is above and below $0.5$, respectively, are

$$|h_1|^2 \alpha \Big/ \left(|h_1|^2 (1-\alpha) + \beta_1\right) \geq \epsilon_1 \tag{18}$$

and

$$|h_1|^2 (1-\alpha) \Big/ \left(|h_1|^2 \alpha + \beta_1\right) \geq \epsilon_2 \tag{19a}$$

$$|h_1|^2 \alpha / \beta_1 \geq \epsilon_1 \tag{19b}$$

Correspondingly, the two success probability expressions are

$$f_1^{Case2}(\alpha) = \begin{cases} \exp\left(-\frac{\zeta_1}{(1+\epsilon_1)\alpha - \epsilon_1} - \frac{\zeta_2}{1-\alpha}\right) & , \alpha > 1 - \frac{1}{1+\epsilon_1} \\ 0 & , \text{otherwise} \end{cases} \tag{20}$$

and

$$f_2^{Case2}(\alpha) = \begin{cases} \exp\left(-\max\left(\frac{\lambda_1 \epsilon_2 \beta_1}{1-(1+\epsilon_2)\alpha}, \frac{\zeta_1}{\alpha}\right) - \frac{\zeta_2}{1-\alpha}\right) \\ \qquad\qquad , \alpha < \frac{1}{1+\epsilon_2} \\ 0 \qquad\qquad , \text{otherwise} \end{cases} \tag{21}$$

respectively. Then, the optimal power allocation is given in Theorem 2.

*Theorem 2:* For case 2, the success probability is maximized when

$$\alpha = \begin{cases} x_1^{Case2} & \text{, if } f_1^{Case1}\left(x_1^{Case2}\right) \leq f_2^{Case2}\left(x_2^{Case2}\right) \\ x_2^{Case2} & \text{, otherwise} \end{cases} \tag{22}$$

where $x_1^{Case2}$ and $x_2^{Case2}$ depend on the ratio $\zeta = \frac{\zeta_1}{\zeta_2} > 1$ in the following way

$$x_1^{Case2} = 1 - \frac{1}{\left(\sqrt{\zeta\left(1+\epsilon_1\right)} + \epsilon_1 + 1\right)} \tag{23}$$

$$x_2^{Case2} = \min\left(\frac{1}{\frac{\epsilon_2}{\epsilon_1} + 1 + \epsilon_2}, 0.5\right) \tag{24}$$

In case 3, each user has cached the desired file of the other, hence, the interference can be eliminated at both users regardless of $\alpha$'s value. In addition, because of no interference, this case is similar to the context of OMA excepting that each user can use the whole spectrum. The success conditions can be written as

$$\left|h_1\right|^2 \alpha / \beta_1 \geq \epsilon_1 \tag{25}$$

$$\left|h_2\right|^2 (1-\alpha) / \beta_2 \geq \epsilon_2 \tag{26}$$

Then, the success probability is

$$f^{Case3}(\alpha) = \exp\left(-\frac{\zeta_1}{\alpha} - \frac{\zeta_2}{1-\alpha}\right) \tag{27}$$

and the optimal power expression is suggested in Theorem 3.

*Theorem 3:* For case 3, the success probability is maximized when

$$\alpha = 1 - \frac{1}{\sqrt{\zeta} + 1} \tag{28}$$

with $\zeta = \frac{\zeta_1}{\zeta_2} > 1$.

Note that the power allocation in Theorem 3 always satisfies $\alpha \geq 0.5$. This implies that if both users can use their cached content to eliminate the interference, the user having a worse condition should be allocated more power. In case 4, the success conditions and the success probability expressions when $\alpha$ is above and below $0.5$, respectively, are

$$\left|h_1\right|^2 \alpha \Big/ \left(\left|h_1\right|^2 (1-\alpha) + \beta_1\right) \geq \epsilon_1 \tag{29a}$$

$$\left|h_2\right|^2 \alpha \Big/ \left(\left|h_2\right|^2 (1-\alpha) + \beta_2\right) \geq \epsilon_1 \tag{29b}$$

$$\left|h_2\right|^2 (1-\alpha) / \beta_2 \geq \epsilon_2 \tag{29c}$$

with success probability function

$$f_1^{Case4}(\alpha) = \begin{cases} \exp\left(-\frac{\zeta_1}{(1+\epsilon_1)\alpha - \epsilon_1} - \max\left(\frac{\lambda_2\epsilon_1\beta_2}{(1+\epsilon_1)\alpha-\epsilon_1}, \frac{\zeta_2}{1-\alpha}\right)\right) \\ \qquad\qquad \text{, } \alpha > 1 - \frac{1}{1+\epsilon_1} \\ 0 \qquad\qquad \text{, otherwise} \end{cases} \tag{30}$$

and

$$\left|h_1\right|^2 (1-\alpha) \Big/ \left(\left|h_1\right|^2 \alpha + \beta_1\right) \geq \epsilon_2 \tag{31a}$$

$$\left|h_1\right|^2 \alpha / \beta_1 \geq \epsilon_1 \tag{31b}$$

$$\left|h_2\right|^2 (1-\alpha) \Big/ \left(\left|h_2\right|^2 \alpha + \beta_2\right) \geq \epsilon_2 \tag{31c}$$

with success probability function

$$f_2^{Case4}(\alpha) = \begin{cases} \exp\left(\max\left(\frac{\lambda_1\epsilon_2\beta_1}{1-(1+\epsilon_2)\alpha}, \frac{\zeta_1}{\alpha}\right) - \frac{\zeta_2}{1-(1+\epsilon_2)\alpha}\right) \\ \qquad\qquad \text{, } \alpha < \frac{1}{1+\epsilon_2} \\ 0 \qquad\qquad \text{, otherwise} \end{cases} \tag{32}$$

The optimal power allocation for this case is given in Theorem 4.

*Theorem 4:* For case 4, the success probability is maximized when

$$\alpha = \begin{cases} x_1^{Case4} & \text{, if } f_1^{Case4}\left(x_1^{Case4}\right) \leq f_2^{Case4}\left(x_2^{Case4}\right) \\ \min\left(x_2^{Case4}, \frac{1}{\frac{\epsilon_2}{\epsilon_1}+\epsilon_2+1}\right) & \text{, otherwise} \end{cases} \tag{33}$$

where $x_1^{Case4}$ and $x_2^{Case4}$ depend on the ratio $\zeta = \frac{\zeta_1}{\zeta_2} > 1$ in the following way

$$x_1^{Case4} = 1 - \frac{1}{\sqrt{\zeta\left(1+\epsilon_1\right)} + \epsilon_1 + 1} \tag{34}$$

$$x_2^{Case4} = \min\left(\frac{1}{1+\epsilon_2}\left(1 - \frac{1}{\sqrt{\zeta\left(1+\epsilon_2\right)}+1}\right), 0.5\right) \tag{35}$$

Generally, in the cache-aided NOMA system, if a user has cached the desired content of other user occupying the same subchannel, the information can be used to remove the interference. Therefore, the QoS is enhanced even without cooperation among users. This can be considered as another type of cache hit which is not commonly considered in the caching literature. On the other hand, caching further contributes to the spectrum and power efficiency improvement target of NOMA by exploiting user's preferences. Therefore, it is a good idea to combine these two advanced techniques in the future networks.

### *B. Multiple user pairs with multiple subchannels*

Previously, we present optimal ways to share a given power $P$ to user in a pair. However, the performance can be further improved by optimizing $P$. Therefore, in this subsection, we present an optimal power allocation policy across user pairs. We denote $P_{\max}$ to be the total consumable power for the transmission, and $P_i$ to be the power allocated to the user pair in the $i$-th subchannel satisfying $\sum_{i=1}^{K/2} P_i = P_{\max}$. Recall that we previously denote $\beta_1 = \frac{d_1^{\gamma}\sigma^2}{P}$ and similarly for $\beta_2$. In this subsection, we will use $P_i$ in place of $P$ to indicate which subchannel (or user pair) is being considered since we are dealing with a multi-subchannel-multi-user scenario.

From the obtained results in the previous subsection, it can be figured out that except for the case when both users are

satisfied by their own caches, in all of the remaining cases, the success probability for a user pair in the $i$-th subchannel will have the form

$$g_i\left(P_i\right) = \exp\left(-\frac{\Psi_i}{P_i}\right) \tag{36}$$

where $\Psi_i, \forall i = 1, \ldots, K/2$ can be obtained by following results in subsection III-A to allocate power to users in each pair. The parameters $\Psi_i$ represents the dependence of the power allocation in this stage on that of the previous stage. Consequently, the success probability of every user in all $K/2$ subchannels is given by

$$\mathcal{G}\left(\mathbf{P}\right) = \exp\left(-\sum_{j=1}^{K/2}\frac{\Psi_j}{P_j}\right) \tag{37}$$

where vector $\mathbf{P}$ contains $P_i, \forall i = 1, \ldots, K/2$. Therefore, maximizing $\mathcal{G}\left(\mathbf{P}\right)$ is equivalent to solving the following convex optimization problem

$$\min_{\mathbf{P}} \sum_{i=1}^{K/2}\frac{\Psi_i}{P_i} \tag{38}$$

$$\text{s.t.} \sum_{i=1}^{K/2} P_i = P_{\max} \tag{39}$$

$$P_i \geq 0, \forall i = 1, \ldots, K/2 \tag{40}$$

whose closed-form solution can be obtained from KKT conditions as follow

$$P_i = \frac{\sqrt{\Psi_i}}{\sum_{j=1}^{K/2}\sqrt{\Psi_j}} P_{\max}, \forall i = 1, \ldots, K/2 \tag{41}$$

In summary, there is a two-stage power allocation presented in this work. Given that each subchannel is occupied by two users, the quantities $\Psi_i, \forall i = 1, \ldots, K/2$ can be defined following results derived in the previous subsection. Then, the total power $P_{\max}$ can be shared to every user pair according to (41).

## IV. Illustrative Results

This section is devoted to present simulation results illustrating the combination of caching and NOMA analyzed above. In this part, we assume the Zipf distribution for the file popularity as has been shown and widely investigated in previous works, i.e. the probability that file $g_i$ is requested by a user is given by $p_i = \left(i^s \sum_{n=1}^{N} \frac{1}{n^s}\right)^{-1}$ where the parameter $s$ corresponds to the skewness and $N$ is the total number of available files in the catalog. This implies that $p_1 > p_2 > \ldots > p_N$. Therefore, in the context of no cooperation between users, the optimal caching policy is to cache from the most to the least popular file. In the experiments corresponding to Fig. 2 and 3, we consider a single subchannel with two users assuming that the power $P$ has been allocated to that pair, and the power coefficient $\alpha$ is defined according to results in subsection III-A. Fig. 4 investigates the whole system with $K = 20$ users and 10 subchannels where $P_{\max}$ is the total power to be shared to every user pair. Default values for some components are given in the below table,

| Parameters | Values |
|---|---|
| Maximum number of file each user can cache | 1 |
| Total number of files | 5 |
| $P$ | 10 |
| $P_{\max}$ | 200 |
| $\sigma^2$ | 1 |
| $(d_1, d_2)$ | $(2, 1)$ |
| $(\lambda_1, \lambda_2)$ | $(2, 1)$ |
| $(\epsilon_1, \epsilon_2)$ | $(1, 1)$ |
| $\gamma$ | 2 |
| $s$ | 0.5 |

and any change will be stated explicitly in each experiment. Besides that, OMA will be considered as a second candidate for the performance comparison. Regarding this, we assume that the spectrum is divided evenly between users, and the power is optimally allocated which is similar to that of NOMA in the above case 3.[2] In addition, when the two users request the same content file, the whole spectrum will be used to broadcast signal to both, hence NOMA and OMA are indifferent in such a situation.

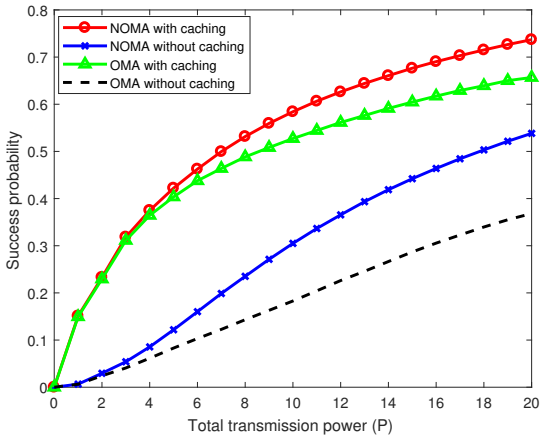


Fig. 2. The improvement in success probability with respect to (w.r.t.) transmission power for both NOMA and OMA schemes with and without caching applied.

Fig. 2 illustrates the variation of our defined success probability w.r.t. the increase in transmission power. The two involved candidates are NOMA and OMA in which NOMA outperforms its counterpart completely. However, the gap between them shrinks when caching is introduced. This is because with caching, NOMA and OMA only differentiate from each other in the last four cases mentioned in subsection III-A. Moreover, these cases occur with low probability when

[2]The expression (28) is optimal for OMA scheme only when the subchannel is divided evenly between two users. This is because the noise power $\sigma^2$ in the definition of $\zeta$ will be cancelled without resulting in any scaling constant making $\alpha$ unchanged. In other words, if the bandwidth given to the first user is $L$ times larger than that of the second user, the first user will suffer from $L$ times more noise power, then $\zeta$ should be replaced by $L\zeta$ in (28) for OMA.

the skewness factor of the Zipf distribution is large. That is to say, if the user preference concentration is high, their requests can be satisfied by their own caches in most of the time without the use of either NOMA or OMA.

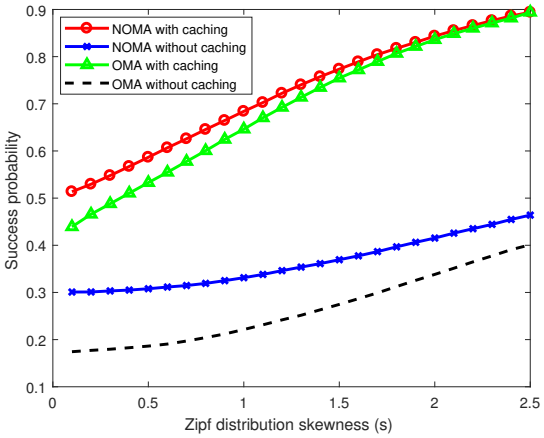


Fig. 3. The effect from the user's preferences on the success probability.

Fig. 3 shows how user's preference is exploited to increase the success probability by NOMA and OMA with and without caching. As $s$ is increased, users will request lower-index files with higher probability. Obviously, as the support of caching, both candidate's performance is picked up remarkably, which emphasizes the benefit of integrating caching into NOMA and OMA. Besides that, with or without caching, the two indicated schemes are closer to each other as the user's preferences concentration increases. Because for larger $s$, there is higher chance for two users to request the same file where NOMA and OMA act the same in this context.

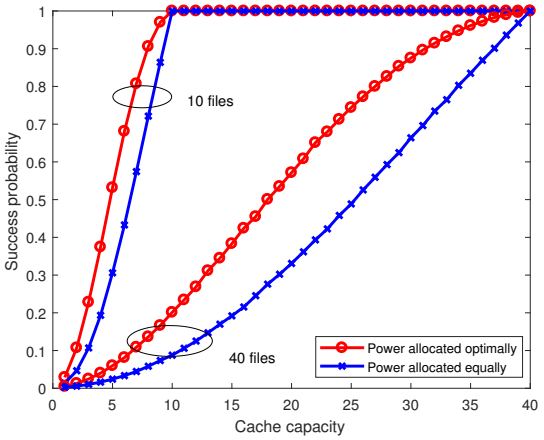


Fig. 4. The influence of the cache capacity enhancement on the multi-user system performance in the small and large file-library-size (10 and 40 files, respectively) scenarios.

Figure 4 illustrates the success probability in a whole system with 20 users where users are paired and assigned different subchannels. The two schemes applied to allocate power to each user pair are the optimal one given in (41) and the flat power allocation scheme where power is split equally between user pairs. The results show that the success probability is improved significantly with the optimal scheme, especially, when the file catalog is large, which approaches the practice.

## V. Conclusion

In this work, we have combined caching and NOMA in a cellular network with Rayleigh fading channels. In order to ensure the fairness among users, we derive an optimal power allocation policy to maximize the success probability in closed form. The analytical and simulation results reveal that NOMA introduces another type of cache hit occuring when a user caches the desired content of the others occupying the same subchannel. This helps users to remove or reduce the interference in the superposed signal even without user cooperation, which increases the success probability. This feature, in addition to the advantages of caching, shows that the combination of these two techniques is a promising idea for future networks.

## References


[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will be 5G?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[3] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.

[4] Y. Shen, C. Jiang, T. Q. S. Quek, and Y. Ren, "Device-to-device-assisted communications in cellular networks: An energy efficient approach in downlink video sharing scenario," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1575–1587, Feb. 2016.

[5] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Fundamental limits of caching in D2D wireless networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 5001–5015, Feb. 2016.

[6] R. Wang, X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[7] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Aug. 2017.

[8] Z. Ding, Z. Wei, J. Yuan, D. W. K. Ng, and M. Elkashlan, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[9] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[10] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744 – 2757, Dec. 2017.

[11] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Tech. Conf.*, Dresden, Germany, Jun. 2013.

[12] F. Fang, H. Zhang, J. Cheng, and V. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722 – 3732, Sep. 2016.

[13] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *CoRR*, vol. abs/1709.06951, 2017. [Online]. Available: http://arxiv.org/abs/1709.06951