Reliability-Oriented Designs in UAV-assisted NOMA Transmission with Finite Blocklength Codes and Content Caching

Yang Yang and M. Cenk Gursoy
Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244
Email: yyang82@syr.edu, mcgursoy@syr.edu

***Abstract*—In this paper, we investigate the reliability in an unmanned aerial vehicle (UAV) assisted caching-based downlink network where non-orthogonal multiple access (NOMA) transmission and finite blocklength (FBL) codes are adopted. In this network, the ground user equipments (GUEs) request contents from a distant base station (BS) but there are no direct links from the BS to the GUEs. A UAV with limited cache size is employed to assist the BS to complete the communication by either first requesting the uncached contents from the BS and then serving the GUEs or directly sending the cached contents to the GUEs. In this setting, we first introduce the decoding error rate in the FBL regime as well as the caching policy at the UAV, and subsequently we construct an optimization problem aiming to minimize the maximum end-to-end decoding error rate among all GUEs under both coding length and maximum UAV transmission power constraints. A two-step alternating algorithm is proposed to solve the problem and numerical results demonstrate that our algorithm can solve the optimization problem efficiently. More specifically, loosening the FBL constraint, enlarging the cache size and having a higher transmission power budget at the UAV lead to an improved performance.**

***Index Terms*—Unmanned aerial vehicle (UAV), non-orthogonal multiple access (NOMA), finite blocklength (FBL) codes, content caching.**

I. Introduction

Recently, unmanned aerial vehicles (UAVs) have been widely deployed for a variety of purposes, including wireless coverage and smart city applications [1, 2]. UAVs are also considered to be promising in numerous 5G applications due to their inherent characteristics, including fast mobility, lower cost, and flexibility in deployment [3]. More specifically, the wireless communication network may leverage low-altitude UAVs by swiftly deploying them and offering significantly improved coverage [4]. These aforementioned benefits indicate that the UAV-enabled communication systems will become increasingly significant in the future wireless networks.

With the rapid development of 5G networks, the data traffic demand in wireless communication has dramatically increased. In most cases, the repeated downloads of a few popular contents are considered as the primary cause of the data traffic congestion. To alleviate this problem, one promising technology is edge caching that enables the edge server to cache popular contents. In certain scenarios, UAVs can operate as an edge server to serve the ground user equipments (GUEs) and cache several popular contents. In [5], the authors have investigated the joint optimization of UAV deployment, caching placement and user association in UAV-assisted cellular networks to maximize the mean opinion score (MOS) of all the users in the cell.

As another promising technology, non-orthogonal multiple access (NOMA) has been thoroughly investigated with relays, and it has been shown to provide outstanding results in improving the performance of overloaded networks [6]. Furthermore, it is also well known that NOMA can improve the spectral efficiency significantly, and thereby has a remarkable potential to enable low-latency communications by serving multiple users simultaneously. It is expected that when NOMA transmission with successive interference cancellation (SIC) at the receiver is combined with the UAV, the wireless propagation environment will be further improved. The performance of NOMA in short-packet communications compared with orthogonal multiple access (OMA) in the finite blocklength (FBL) regime has been investigated in [7]. The authors in [8] have maximized the sum rate by optimally determining the position of UAV and the power allocation in NOMA.

Ultra-reliable and low latency communication (URLLC) is one of the use cases in 5G networks to address the mission-critical services [9]. In URLLC, short packets with FBL codes have been utilized to decrease the transmission delay. As a consequence, wireless communication system design and performance analysis should be dramatically modified. More explicitly, the traditional Shannon's information capacity is no longer applicable in the FBL regime. Therefore, the decoding error probability cannot be neglected. The authors in [10] have derived an accurate approximation of the transmission rate with FBL codes for the additive white Gaussian noise (AWGN) channel, and the decoding error probability has been explicitly investigated. In [11], the authors have analyzed the global optimal resource allocation for URLLC in the FBL regime.

Among existing works, the authors in [12] have studied the UAV-assisted downlink transmission model by considering the two-user NOMA case under constraints on energy and the caching capacity at the UAV. The authors in [13] have investigated UAV deployment and content placement in a cache-enabled multi-UAV network to minimize the average request delay of users. From another aspect, the achievable effective capacity comparison between two-user NOMA and its OMA counterpart under delay quality-of-service constraints in the FBL regime has been investigated in [14].

In this paper, we merge FBL regime with NOMA and content caching in a UAV-assisted network, aiming to minimize the maximum end-to-end decoding error probability when multiple GUEs are considered. In contrast to previous works, we consider the scenario that may involve more than 2 GUEs and combine the content caching with FBL regime so that the data traffic burden is alleviated, and correspondingly the performance is enhanced. Our main contributions in this paper are summarized as follows:

1) We describe and analyze the UAV-assisted downlink NOMA tranmissions with FBL codes and content caching.
2) We investigate the end-to-end decoding error probability at the GUE and propose a caching policy at the UAV.
3) We develop a two-step alternating optimization algorithm to minimize the maximum end-to-end decoding error rate among all GUEs under both coding length and maximum UAV transmission power constraints.

The remainder of this paper is organized as follows. In Section II, we first introduce the system model and describe the FBL regime as well as the SINR when NOMA transmission is utilized, followed by the explicit specification of the end-to-end decoding error probability and the caching policy. In Section III, we first construct an optimization problem aiming to minimize the maximum end-to-end decoding error rate among all GUEs under both coding length and maximum UAV transmission power constraints, and then we propose a two-step alternating optimization algorithm to address the problem. Simulation results are provided in Section IV. Finally, in Section V, we draw conclusions.

II. System Model

In this paper, we investigate a downlink system model where a BS, a UAV and a set $\mathcal{N} = \{1, 2, ..., N\}$ of N GUEs are considered, as shown in Fig.1. Each terminal has a single antenna. Due to the blockages, e.g., as a result of natural terrains or large cluster of buildings., we presume that there are no available direct links from the BS to the GUEs. Therefore, a UAV with limited cache size is deployed to serve the GUEs by utilizing NOMA transmission with FBL codes. Such a UAV is set to operate at a fixed height, and it can move on the same plane. All channels are assumed to be quasi-static/unchanged within a transmission frame, and thereby we consider that the impact of UAV movement is limited, i.e., the UAV position is fixed during a frame. More explicitly, the optimized parameters, e.g., transmission power allocations at the UAV, are effective in the current transmission frame.

Let C_{uav} denote the cache size at the UAV, and we assume there are a total of C contents that can be requested by the GUEs and the size of the c-th content is set to be I_c bits. If the requested content can be found in the cache of the UAV, then it will be transmitted to the GUE directly without involving the BS, otherwise the UAV will request this content from the BS first and then start transmitting.

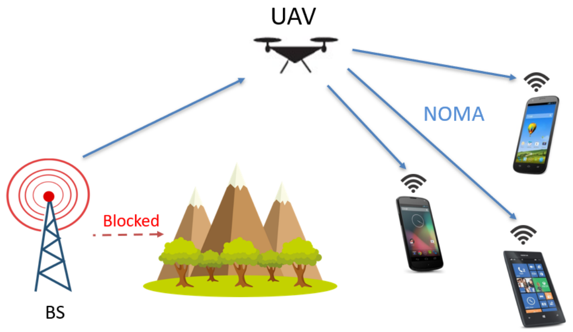


Fig. 1: An illustration of the considered network.

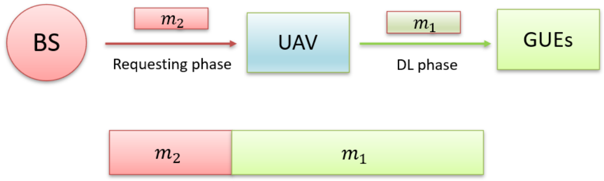


Fig. 2: System topology and frame structure.

A. FBL Transmission with Caching

The duration of a transmission symbol is denoted by T_{syb}. With this, the delay limitation of T in seconds corresponds to $M = T/T_{\text{syb}}$ symbols. More specifically, T seconds or equivalently M symbol durations serve as a bound on the frame length of the service completion of the requested content/task. A requesting phase with a length of m_2 symbols and a downlink (DL) transmission phase with a length of m_1 symbols are the two phases in a frame, as depicted in Fig. 2. In this paper, we introduce $X_{c,n,i} \in \{0, 1\}$ to indicate the request of the n-th GUE ($X_{c,n,i} = 1$ if the n-th GUE is requesting content c in the i-th frame). The size of the requested content for the n-th GUE in the i-th frame is $D_{n,i} = \sum_{c=1}^{C} X_{c,n,i} I_c$ bits. Note that within each frame, each GUE can only request one content, e.g., $\sum_{c=1}^{C} X_{c,n,i} = 1, \forall n \in \mathcal{N}$. The UAV will first check its cache: if the requested content has been cached, then there is no need to consult the BS, otherwise such content is required to be downloaded from the BS. After the UAV checks its cache for all the requested contents in the i-th frame, the UAV will download all the uncached but requested contents from the BS via a wireless link in the requesting phase with a duration of $m_2 T_{\text{syb}}$ seconds. Then, in the DL transmission phase, whose duration is $m_1 T_{\text{syb}}$ seconds, the UAV will send all the requested contents to the GUEs via NOMA transmissions. It is obvious that the total service time of every content request is constrained by $m_1 + m_2 = M$. Following [10], the coding rate R in the FBL regime is approximated as

$$R \approx \log_2(1+\gamma) - \sqrt{\frac{V}{m}}\frac{Q^{-1}(\varepsilon)}{\ln 2}, \tag{1}$$

where ε is the decoding error probability, m is the blocklength, γ is the signal-to-noise ratio (SNR)/signal-to-interference-plus-noise ratio (SINR) at the receiver, Q^{-1} is the inverse function of $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}dt$ and V is the channel dispersion defined as $V = 1-(1+\gamma)^{-2}$.

In this paper, we define $Y_{c,i} \in \{0,1\}$ to be the caching indicator: if $Y_{c,i} = 1$, it is indicated that content c has been cached at the UAV in the i-th frame. We further define $Z_{c,i}$ as the requesting indicator, as follows:

$$Z_{c,i} = \begin{cases} 1 & \text{when } \sum_{n=1}^{N} X_{c,n,i} \geq 1; \\ 0 & \text{when } \sum_{n=1}^{N} X_{c,n,i} = 0, \end{cases} \tag{2}$$

with $Z_{c,i} = 1$ indicating that content c is requested in the i-th frame by one or multiple GUEs. Consequently, in the i-th frame, the total size of requested but uncached contents is $D_{\text{uav,i}} = \sum_{c=1}^{C} Z_{c,i}(1-Y_{c,i})I_c$ bits. Since the target coding rate in the requesting phase is $R_{\text{uav,i}} = \frac{D_{\text{uav,i}}}{m_2}$, the decoding error probability of the UAV in the i-th frame in the requesting phase can be expressed as

$$\varepsilon_i^{\text{UAV}} \approx Q\left(\sqrt{\frac{m_2}{V_{\text{uav,i}}}}\left(\log_2(1+\gamma_{\text{uav,i}}) - \frac{D_{\text{uav,i}}}{m_2}\right)\log_e 2\right). \tag{3}$$

Considering $R_{n,i} = \frac{D_{n,i}}{m_1}$ as the target achievable coding rate of the n-th GUE in the i-th frame, the decoding error probability in the DL phase can be expressed as

$$\varepsilon_{n,i} \approx Q\left(\sqrt{\frac{m_1}{V_{n,i}}}\left(\log_2(1+\gamma_{n,i}) - \frac{D_{n,i}}{m_1}\right)\log_e 2\right). \tag{4}$$

Note that since we operate in the FBL regime, the blocklength of each frame is limited by M and the decoding error probability at the receiver is non-negligible.

B. SINR in Transmissions

Based on (3) and (4), we know that SINR can affect the decoding error probability significantly, and hence in this section we explicitly introduce the SINR in different transmissions.

In the requesting phase, the UAV is downloading data from the BS. Since all channels are assumed to be quasi-static, we consider that the channels remain constant within a frame. Therefore, the SNR for the UAV in the requesting phase in the i-th frame is given by

$$\gamma_{\text{uav,i}} = \rho_{\text{uav}}|h_{\text{uav,i}}|^2, \tag{5}$$

where $\rho_{\text{uav}} = \frac{P_{\text{BS}}}{\sigma^2}$, $h_{\text{uav,i}}$ is the channel coefficient between the UAV and the BS, P_{BS} is the transmission power from the BS to the UAV and σ^2 denotes the power of the AWGN.

In the DL phase, the UAV broadcasts the superposed signals to all GUEs in accordance with the NOMA principle. As a result, the received signal at each GUE in the i-th frame is expressed as

$$y_{n,i} = h_{n,i}\sum_{k=1}^{N}\sqrt{P_{\max}\rho_{k,i}}x_{k,i} + \eta, \forall n \in \mathcal{N}, \tag{6}$$

where $x_{k,i}$, $\rho_{k,i}$ denote the message and the power allocation factor of the k-th GUE in the i-th frame, respectively. $P_{\max}$ is the transmission power constraint/budget at the UAV, η represents the AWGN, e.g., $\eta \sim \mathcal{CN}(0,\sigma^2)$, and $h_{n,i}$ is channel coefficient between the UAV and the n-th GUE in the i-th frame. Note that $\sum_{k=1}^{N}\rho_{k,i} = 1$.

In order to implement the successive interference cancellation (SIC) in NOMA technique, we reorder all the GUEs based on their channel quality at the beginning of each frame. In the i-th frame, all N GUEs are sorted in an increasing order, i.e., $|h_{1,i}| \leq |h_{2,i}| \leq ..., \leq |h_{N,i}|$. The GUE that has the worst channel is considered as the first GUE and the last GUE has the best channel. Based on the SIC principle, the n-th ($1 \leq n \leq N$) GUE must first decode the signals of all the previous $n-1$ GUEs, and then those signals are removed from the superposed received signal. Consequently, the SINR for the n-th GUE in decoding its own signal in the i-th frame is described as follows:

$$\gamma_{n,i} = \frac{|h_{n,i}|^2 P_{\max}\rho_{n,i}}{\sum_{t=n+1}^{N}|h_{n,i}|^2 P_{\max}\rho_{t,i} + \sigma^2} \tag{7}$$

In the FBL regime, the SIC errors are non-negligible since the n-th GUE needs to first decode the previous $n-1$ GUEs' signals and then decode its own signal. If SIC is not successful, its own decoding will fail as well. Hence, determining the error rate in decoding the signals of other GUEs is of great importance. The SINR for the n-th user in decoding the k-th ($k \leq n-1 \leq N$) GUE's signal in the i-th frame is formulated as

$$\gamma_{n,k,i} = \frac{|h_{n,i}|^2 P_{\max}\rho_{k,i}}{\sum_{t=k+1}^{N}|h_{n,i}|^2 P_{\max}\rho_{t,i} + \sigma^2} \tag{8}$$

The first GUE can directly decode its own signal by considering the signals of all other GUEs as interference since no SIC is performed at GUE 1. On the other hand, the last

GUE performs SIC of all other GUEs' signals and its SINR becomes quite simple if all SICs are successful:

$$\gamma_{N,i} = \frac{|h_{N,i}|^2 P_{\max} \rho_{N,i}}{\sigma^2}. \tag{9}$$

C. End-to-end Decoding Error Probability

Our objective in this paper is to minimize the maximum end-to-end decoding error rate among all GUEs under both coding length and maximum UAV transmission power constraints. In this section, we analyze the end-to-end decoding error probability for GUEs. For the n-th GUE in the i-th frame, we consider two different scenarios: the requested content has been cached at the UAV or it is uncached.

In the first case, the requested content of the n-th GUE has been cached at the UAV, and the end-to-end decoding error probability $\epsilon_{n,i}^{\text{CA}}$ consists of two components: error probability $\epsilon_{n,k,i}^{\text{SIC}}$ in decoding signals of other GUEs in adopting SIC, and the error probability $\epsilon_{n,i}$ in decoding its own signal. We express $\epsilon_{n,i}^{\text{CA}}$ as follows:

$$\begin{aligned} \epsilon_{n,i}^{\text{CA}} &= 1 - \prod_{k=1}^{n-1} (1 - \epsilon_{n,k,i}^{\text{SIC}})(1 - \epsilon_{n,i}) \\ &\overset{(a)}{\approx} \sum_{k=1}^{n-1} \epsilon_{n,k,i}^{\text{SIC}} + \epsilon_{n,i}. \end{aligned} \tag{10}$$

Here approximation (a) holds since the decoding error probabilities are typically of the order of 10^{-5} in an ultra-reliable communication scenario, and hence all the terms including two or more errors being multiplied can be neglected.

We then investigate the second case in which there are three components in the end-to-end decoding error probability $\epsilon_{n,i}^{\text{UN}}$ of the n-th GUE in the i-th frame: error probability ϵ_i^{UAV} in decoding the downloaded contents from the BS at the UAV, error probability $\epsilon_{n,k,i}^{\text{SIC}}$ in decoding signals of other GUEs in adopting SIC, and the error probability $\epsilon_{n,i}$ in decoding its own signal. In this case, we have

$$\begin{aligned} \epsilon_{n,i}^{\text{UN}} &= 1 - (1 - \epsilon_i^{\text{UAV}}) \prod_{k=1}^{n-1} (1 - \epsilon_{n,k,i}^{\text{SIC}})(1 - \epsilon_{n,i}) \\ &\overset{(b)}{\approx} \epsilon_i^{\text{UAV}} + \sum_{k=1}^{n-1} \epsilon_{n,k,i}^{\text{SIC}} + \epsilon_{n,i}. \end{aligned} \tag{11}$$

Here approximation (b) holds due to the same reason as in approximation (a). Combining these two cases, we can further describe the end-to-end decoding error rate of the n-th GUE in the i-th frame as $\epsilon_{n,i}^{\text{tot}}$:

$$\epsilon_{n,i}^{\text{tot}} = \sum_{c=1}^{C} X_{c,n,i}(1 - Y_{c,i})\epsilon_i^{\text{UAV}} + \sum_{k=1}^{n-1} \epsilon_{n,k,i}^{\text{SIC}} + \epsilon_{n,i}. \tag{12}$$

In (12), ϵ_i^{UAV} can be computed from (3) and $\epsilon_{n,i}$ can be obtained via (4). As for $\epsilon_{n,k,i}^{\text{SIC}}$, it can be calculated as follows:

$$\epsilon_{n,k,i}^{\text{SIC}} \approx Q\left(\sqrt{\frac{m_1}{V_{n,k,i}}}\left(\log_2(1+\gamma_{n,k,i}) - \frac{D_{k,i}}{m_1}\right)\log_e 2\right), \tag{13}$$

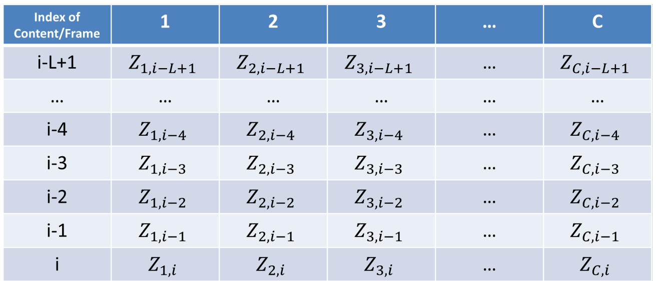
where $\gamma_{n,k,i}$ can be obtained from (8) and $V_{n,k,i} = 1 - (1 + \gamma_{n,k,i})^{-2}$.

D. Caching Policy

In this section, we introduce our caching policy at the UAV. Our main purpose in the caching procedure is to cache the most popular/frequently requested contents. A caching list that stores all the request information of the past L frames is built at the UAV. Before the $i+1$-th frame starts, the UAV will delete the request information of the $i-L$-th frame and put the request information of the i-th frame into the caching list, as shown in Fig. 3. Then the UAV will calculate the popularity of each content. Let $O_{c,i}$ denote the popularity of content c in the i-th frame, which is given by

$$O_{c,i} = \frac{\sum_{l=1}^{L} Z_{c,i-l+1}}{L}. \tag{14}$$

When the popularity of all the contents has been computed, the UAV will cache contents from the highest popularity to the lowest until it reaches the cache size limitation C_{uav}, and then the UAV will update the caching indicator $\{Y\}$, which will be used in the $i+1$-th frame. Fig. 4 illustrates an example of caching list with $i = 50$, $L = 10$ and $C = 5$. If all contents have the same size and the cache size at the UAV only allows to cache 2 contents, the UAV will cache content 1 and content 5 in its cache at the end of the 50th frame.



Index of Content/Frame	1	2	3	...	C
i-L+1	$Z_{1,i-L+1}$	$Z_{2,i-L+1}$	$Z_{3,i-L+1}$	...	$Z_{C,i-L+1}$
...	...	...	...	...	...
i-4	$Z_{1,i-4}$	$Z_{2,i-4}$	$Z_{3,i-4}$	...	$Z_{C,i-4}$
i-3	$Z_{1,i-3}$	$Z_{2,i-3}$	$Z_{3,i-3}$	...	$Z_{C,i-3}$
i-2	$Z_{1,i-2}$	$Z_{2,i-2}$	$Z_{3,i-2}$	...	$Z_{C,i-2}$
i-1	$Z_{1,i-1}$	$Z_{2,i-1}$	$Z_{3,i-1}$	...	$Z_{C,i-1}$
i	$Z_{1,i}$	$Z_{2,i}$	$Z_{3,i}$	...	$Z_{C,i}$

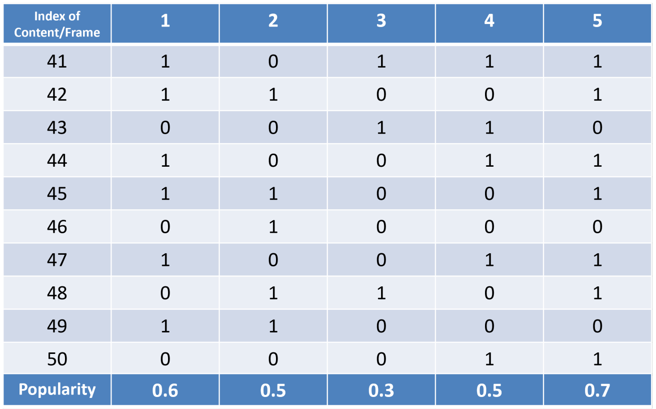
Fig. 3: An illustration of the caching list.

III. Minimization of Maximum Error Probability

In this section, we first formulate and analyze the global maximum error rate minimization problem within a given frame in the considered network and then propose a two-step alternating algorithm to tackle the optimization problem.

A. Problem Formulation

In this paper, our objective is to minimize the maximum end-to-end decoding error rate among all GUEs by jointly



Index of Content/Frame	1	2	3	4	5
41	1	0	1	1	1
42	1	1	0	0	1
43	0	0	1	1	0
44	1	0	0	1	1
45	1	1	0	0	1
46	0	1	0	0	0
47	1	0	0	1	1
48	0	1	1	0	1
49	1	1	0	0	0
50	0	0	0	1	1
Popularity	0.6	0.5	0.3	0.5	0.7

Fig. 4: An example of the caching list with popularity.

determining the GUEs' transmission power allocation factors $\{\rho_n\}$ and the length of the DL phase m_1 subject to the coding length and UAV transmission power constraints. Consequently, in the i-th frame the optimization problem is formulated as follows:

$$\textbf{P1:}\quad \underset{\{\rho_{n,i}\},m_{1,i}}{\textbf{Minimize}}\ \max_{\forall n\in\mathcal{N}}\{\epsilon^{\text{tot}}_{n,i}\} \tag{15}$$

$$\textbf{s. t.}\quad \sum_{n=1}^{N}\rho_{n,i}=1, \tag{15a}$$

$$m_{1,i}+m_{2,i}=M, \tag{15b}$$

$$m_{1,i},m_{2,i}\in\mathbb{Z}. \tag{15c}$$

In **P1**, (15a) is the UAV transmission power constraint and (15b) is the maximum coding length constraint. Solving the non-convex problem **P1** directly is quite challenging due to the strongly coupled parameters $\{\rho_{n,i}\}$, $m_{1,i}$ and highly non-linear objective function. In order to address this, we propose a two-step alternating optimization method that decouples the optimization variables and iteratively solves the problem.

B. Two-step Alternating Optimization

In the j-th optimization iteration during the i-th frame, we first fix $m_{1,i}$ as $m_{1,i,j-1}$ by adopting the optimization results in the $j-1$-th iteration and design the GUEs' transmission power allocation factors $\{\rho_{n,i,j}\}$ in order to decouple the optimization variables. Then, with given $\{\rho_{n,i,j}\}$, we can optimally obtain $m_{1,i,j}$ in the second step, and thereby we use the obtained $\{\rho_{n,i,j}\}$ and $m_{1,i,j}$ in the $j+1$-th iteration.

1) Optimization of GUEs' Transmission Power Allocation Factors: In the j-th iteration, when m_1 is fixed, it is obvious that $m_2 = M - m_1$ is also fixed, and hence we now seek to find the optimal power allocation strategies $\{\rho_{n,i,j}\}$ at the UAV to minimize the maximum end-to-end decoding error rate among all GUEs. Therefore, **P1** is transformed into **P2** when m_1 is fixed:

$$\textbf{P2:}\quad \underset{\{\rho_{n,i,j}\}}{\textbf{Minimize}}\ \max_{\forall n\in\mathcal{N}}\{\epsilon^{\text{tot}}_{n,i,j}\} \tag{16}$$

$$\textbf{s. t.}\quad \sum_{n=1}^{N}\rho_{n,i,j}=1, \tag{16a}$$

where $\{\rho_{n,i,j}\}$ and $\epsilon^{\text{tot}}_{n,i,j}$ are the power allocation factors at the UAV and the end-to-end decoding error probability of the n-th GUE in the j-th optimization iteration during the i-th frame, respectively.

P2 is still a min-max optimization problem, and it is nontrivial to solve. In order to address the min-max problem, we further transform **P2** into N sub-problems:

$$\textbf{P2A:}\quad \underset{\{\rho_{n,i,j}\}}{\textbf{Minimize}}\ \epsilon^{\text{tot}}_{n,i,j} \tag{17}$$

$$\textbf{s. t.}\quad \sum_{n=1}^{N}\rho_{n,i,j}=1, \tag{17a}$$

$$\epsilon^{\text{tot}}_{n,i,j}\geq\epsilon^{\text{tot}}_{k,i,j},\qquad \forall k\neq n\in\mathcal{N} \tag{17b}$$

For each GUE $n\in\mathcal{N}$, we construct a sub-problem. In each sub-problem **P2A**, we only minimize the end-to-end decoding error probability for a single GUE and (17b) assures such minimized error probability is the maximal one among all GUEs, and hence the obtained power allocation strategies $\{\rho_{n,i,j}\}$ might be one solution of **P2A**. We then introduce **Lemma 1** to attain the solution of **P2** from **P2A**.

Lemma 1: *Among all the sub-problems **P2A**, the one which achieves the minimum value in the objective function has the same solution as **P2**.*

Proof: Suppose that the t-th sub-problem achieves the minimum value of the objective function, e.g., $\epsilon^{\text{tot}*}_{t,i,j}<\epsilon^{\text{tot}*}_{v,i,j},\forall v\neq t\in\mathcal{N}$. When we substitute the solution of the t-th sub-problem into **P2**, the value of the objective function should be the same as $\epsilon^{\text{tot}*}_{t,i,j}$.

If the solution of **P2** is not the same as the t-th sub-problem, i.e., $\epsilon^{\text{tot}*}_{u,i,j},u\neq t$ is the minimum achievable error probability with the solution of **P2**, we should have $\epsilon^{\text{tot}*}_{u,i,j}<\epsilon^{\text{tot}*}_{t,i,j}$ since **P2** is a minimization problem, and the solution leading to $\epsilon^{\text{tot}*}_{u,i,j}$ in **P2** must be the solution of the u-th sub-problem. However, according to our initial assumption, we should also have $\epsilon^{\text{tot}*}_{t,i,j}<\epsilon^{\text{tot}*}_{u,i,j}$, which is contrary to $\epsilon^{\text{tot}*}_{u,i,j}<\epsilon^{\text{tot}*}_{t,i,j}$, and hence the solution of **P2** must be the same as the t-th sub-problem which achieves the minimum value of the objective function among all N sub-problems. □

Combining the solutions from all the N sub-problems, based on **Lemma 1**, we know the one that provides us the minimum end-to-end decoding error probability in the objective function is the solution of **P2**.

Each sub-problem **P2A** can be solved via a nonlinear optimization tool. However, the Q function increases the computational complexity dramatically. To tackle this, following the approach in [15], we can approximate the Q function by the following function F with any fixed m and D, e.g.,

$Q(\gamma, m, D) \approx F_m^D(\gamma)$:

$$F_m^D(\gamma) = \begin{cases} 1, & \gamma \leq \theta_m^D \\ \frac{1}{2} - \alpha_m^D(\gamma - \beta_m^D), & \theta_m^D < \gamma < \kappa_m^D \\ 0, & \gamma \geq \kappa_m^D \end{cases} \tag{18}$$

where $\alpha_m^D = \sqrt{\frac{m}{2\pi 2^{\frac{2D}{m}} - 1}}$, $\beta_m^D = 2^{\frac{2D}{m}} - 1$, $\theta_m^D = \beta_m^D - \frac{1}{2\alpha_m^D}$ and $\kappa_m^D = \beta_m^D + \frac{1}{2\alpha_m^D}$.

By employing (18), with the given m and D, the overall end-to-end decoding error rate of the n-th GUE in the i-th frame becomes $\epsilon_{n,i}^{\mathrm{F}}$:

$$\begin{aligned} \epsilon_{n,i}^{\mathrm{tot}} \approx \epsilon_{n,i}^{\mathrm{F}} = & \sum_{c=1}^{C} X_{c,n,i}(1 - Y_{c,i}) F_{m_{2,i}}^{D_{\mathrm{uav,i}}}(\gamma_{\mathrm{uav,i}}) + \\ & \sum_{k=1}^{n-1} F_{m_{1,i}}^{D_{k,i}}(\gamma_{n,k,i}) + F_{m_{1,i}}^{D_{n,i}}(\gamma_{n,i}). \end{aligned} \tag{19}$$

We can then transform **P2A** into **P2B**:

$$\textbf{P2B:} \quad \underset{\{\rho_{n,i,j}\}}{\textbf{Minimize}} \ \epsilon_{n,i,j}^{\mathrm{F}} \tag{20}$$

$$\textbf{s. t.} \quad \sum_{n=1}^{N} \rho_{n,i,j} = 1, \tag{20a}$$

$$\epsilon_{n,i,j}^{\mathrm{F}} \geq \epsilon_{v,i,j}^{\mathrm{F}}, \quad \forall v \in \mathcal{N}, v \neq n \tag{20b}$$

$$\theta_{m_{1,i,j}}^{D_{n,i,j}} < \gamma_{n,i,j} < \kappa_{m_{1,i,j}}^{D_{n,i,j}}, \tag{20c}$$

$$\theta_{m_{1,i,j}}^{D_{k,i,j}} < \gamma_{n,k,i,j} < \kappa_{m_{1,i,j}}^{D_{k,i,j}}, \quad \forall k \in \mathcal{N}, k \leq n-1 \tag{20d}$$

where $\epsilon_{n,i,j}^{\mathrm{F}}$ is $\epsilon_{n,i}^{\mathrm{F}}$ in the j-th optimization iteration during the i-th frame. **P2B** can still be solved via a nonlinear optimization tool with no Q function in it, which reduces the computational complexity at the expense of lower accuracy due to the approximation. One should balance the solution accuracy and the computational complexity in choosing **P2A** or **P2B**.

By solving either **P2A** or **P2B**, we can obtain the optimal power allocation factors $\{\rho_n\}^*$ at the UAV. Such obtained $\{\rho_n\}^*$ in the j-th iteration during the i-th frame is denoted as $\{\rho_{n,i,j}\}$.

2) Optimization of the Duration of DL Phase: In the second step of the two-step alternating optimization algorithm, we fix the power allocation factors $\{\rho_n\}$ at the UAV to be $\{\rho_{n,i,j}\}$, and then **P1** becomes **P3** to obtain the optimal duration/symbol length of the DL phase in the j-th iteration during the i-th frame:

$$\textbf{P3:} \quad \underset{m_{1,i,j}}{\textbf{Minimize}} \ \max_{\forall n \in \mathcal{N}} \{\epsilon_{n,i,j}^{\mathrm{tot}}\} \tag{21}$$

$$\textbf{s. t.} \quad m_{1,i,j} + m_{2,i,j} = M, \tag{21a}$$

$$m_{1,i,j}, m_{2,i,j} \in \mathbb{Z}, \tag{21b}$$

where $m_{1,i,j}$ and $m_{2,i,j}$ are the symbol lengths of the DL phase and requesting phase in the j-th optimization iteration within the i-th frame, respectively.

P3 is a discrete optimization problem and exhaustive search can be used to obtain the proper $m_{1,i,j}$. However, when M becomes large, exhaustive search will be prohibitive. To address this, we can first relax $m_{1,i,j}$ to be continuous valued and then solve **P3** without (21b) via a nonlinear optimization tool. As for **P2**, similar approach can be utilized for **P3** to transform it into several minimization sub-problems. At the end of the two-step alternating algorithm, we choose the closest integer to the continuous solution as the optimal $m_{1,i}$.

By iteratively solving **P2A**/**P2B** and **P3**, we can obtain the solution of **P1** once they converge. Algorithm 1 below provides a description of the proposed two-step alternating optimization algorithm.

Algorithm 1 Two-step alternating optimization algorithm

```
Initialization:
    1) Initialize $\{\rho_{n,i,0}\}$, $m_{1,i,0}$.
Actions:
    1) For $j = 1 : \mathrm{J}_{\max}$
    2) Obtain $\{\rho_{n,i,j}\}$ by solving P2A/P2B with $m_{1,i,j-1}$.
    3) Obtain $m_{1,i,j}$ by solving P3 with $\{\rho_{n,i,j}\}$.
    4) End If converged.
```

Note that in the last iteration, we need to process action 2) one more time to obtain the final power allocation factors $\{\rho_{n,i}\}$ at the UAV for the i-th frame.

C. Operation Framework in the UAV-assisted Network

In this section, we explicitly describe the overall framework in the considered UAV-assisted downlink network. The detailed framework is illustrated in Algorithm 2 below.

Algorithm 2 Framework in the UAV-assisted Network

```
Initialization:
    1) Initialize the maximum number of frames $\mathrm{I}_{\max}$,
       caching size limitation $C_{\mathrm{uav}}$ at the UAV, total length
       of a frame $M$, transmission power $P_{\mathrm{BS}}$ from the BS
       to the UAV during the requesting phase, maximum
       available transmission power $P_{\max}$ at the UAV during the DL phase.
Actions:
    1) For $i = 1 : \mathrm{I}_{\max}$
    2) Obtain the locations of UAV and GUEs, calculate
       the channel coefficients $h_{\mathrm{uav,i}}$ and $\{h_{n,i}\}$.
    3) Reorder the GUEs in an increasing order, i.e.,
       $|h_{1,i}| \leq |h_{2,i}| \leq ..., \leq |h_{N,i}|$.
    4) Check all the content requests from the GUEs with
       the cached contents at the UAV, generate $\{X_{c,n,i}\}$,
       $\{Y_{c,i}\}$ and $\{Z_{c,i}\}$.
    5) Construct P1 with the parameters in the $i$-th frame.
    6) Solve P1 via Algorithm 1 to obtain the transmission
       power allocation factors $\{\rho_{n,i}\}$ at the UAV in the
```

DL phase during the i-th frame, and the length of the DL phase in the i-th frame $m_{1,i}$.

7) Update the caching list at the UAV, calculate the popularity of each content $\{O_{c,i}\}$, and then update the cache.

8) **End**.

IV. Numerical Results

In this section, we conduct numerical analysis of the minimized maximum end-to-end decoding error probability among all GUEs via the proposed algorithm. We first demonstrate the average min-max end-to-end decoding error rate versus the coding length constraint under different $P_{\max}$ constraints. We then investigate the impact of cache size limitation as well as the length of the caching list at the UAV.

In the simulations, the channels are generated by $h_n = \sqrt{\xi_0 d_n^{-\alpha_n}}\widetilde{g}_n$, $n \in \mathcal{N}$ and $h_{\text{uav}} = \sqrt{\xi_0 d_{\text{uav}}^{-\alpha_{\text{uav}}}}\widetilde{g}_{\text{uav}}$ where $\xi_0 = -30$ dB denotes the path loss at the reference point $d_0 = 1$ m. d_n, α_n and $\widetilde{g}_n$ denote the distance from the UAV to the n-th GUE, path loss exponent, and complex Gaussian distributed fading component for the n-th GUE, respectively. Similarly, d_{uav}, α_{uav}, $\widetilde{g}_{\text{uav}}$ are the distance from the BS to the UAV, path loss exponent, and complex Gaussian distributed fading component of this link. The path loss exponents are set to be $\alpha_n = 3.5$ and $\alpha_{\text{uav}} = 2$. The other parameters are set to be $\sigma^2 = -95$ dBm, $P_{\text{BS}} = 2$W during the requesting phase.

In Fig. 5, we analyze the average min-max error probability attained with the proposed algorithm for 3 GUEs. In Fig. 5, different dashed lines plot the average min-max error rates under different transmission power budgets $P_{\max}$ at the UAV. We observe that the min-max error probability is reduced as the blocklength constraint M increases, which is expected since increasing M is the same as extending the transmission time, resulting in less strict requirements in the coding rate. We further observe that increasing the transmission power budget $P_{\max}$ at the UAV improves the performance as well. By increasing $P_{\max}$, we are more capable to obtain a higher SNR/SINR and hence improve the min-max error probability.

We next analyze the influence of the cache size limitation as well as the length of the caching list in Fig. 6. In Fig. 6, the curves of average min-max error probability versus the cache size limitation C_{uav} at the UAV are plotted with different lengths L of the caching list. From Fig. 6, we observe that the one with a larger caching list always outperforms the other one. With a larger caching list, it is more likely to figure out all the popular contents, resulting in an improved efficiency in the caching procedure. We also observe that larger cache size leads to a lower min-max error rate, which is due to the enhancement in caching at the UAV so that more contents are served without resorting to the BS, and thereby improving the min-max error probability. Note that the rate of improvement becomes smaller as the cache size at the UAV grows beyond approximately 15. This is because the most popular contents are to be cached first and continuously increasing the cache size starts enabling the UAV to cache the contents with less and less popularity, resulting in limited improvement in the min-max error rate. Besides, even if all contents are cached at the UAV, there will still be decoding errors during the DL phase, and hence the min-max end-to-end decoding error probability would not vanish by simply increasing the cache size at the UAV.

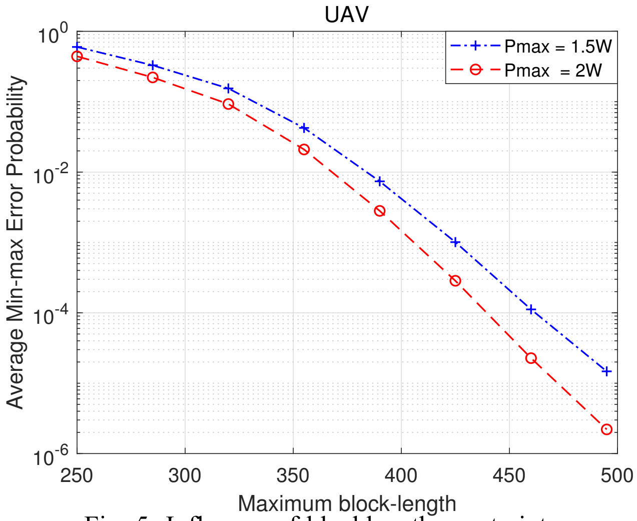


Fig. 5: Influence of blocklength constraint.

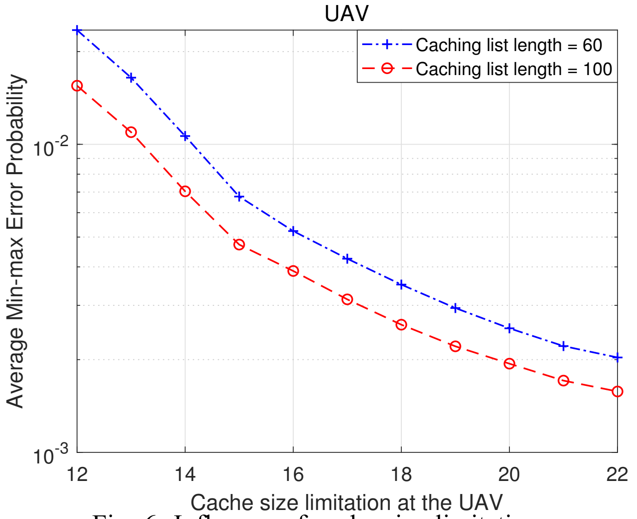


Fig. 6: Influence of cache size limitation.

V. Conclusion

In this paper, we have investigated the reliability in a UAV-assisted caching-based downlink network where NOMA transmission and FBL codes are adopted. We have first introduced the system model and described the FBL regime as well as the SINR when NOMA transmission is utilized. We then identified the end-to-end decoding error probability and specified the caching policy at the UAV. An optimization problem aiming to minimize the maximum end-to-end decoding error rate among all GUEs under both coding length and maximum UAV transmission power constraints has been constructed. Subsequently, we propose a two-step alternating optimization algorithm to solve the problem through which the transmission power allocation factors at the UAV and the length of the DL phase are optimally determined. Numerical results demonstrate that the higher power budget $P_{\max}$ is at the UAV, the smaller end-to-end decoding error rate is expectedly attained, and a larger maximum blocklength M leads to an improved performance in the network. We have further observed that content caching at the UAV can significantly improve the end-to-end decoding error probability.

References

[1] X. Zhang and L. Duan, "Fast deployment of UAV networks for optimal wireless coverage," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 588–601, 2018.

[2] F. Mohammed, A. Idries, N. Mohamed, J. Al-Jaroodi, and I. Jawhar, "UAVs for smart cities: Opportunities and challenges," in *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 267–273, IEEE, 2014.

[3] Z. Ullah, F. Al-Turjman, and L. Mostarda, "Cognition in UAV-aided 5G and beyond communications: A survey," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 872–891, 2020.

[4] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications magazine*, vol. 54, no. 5, pp. 36–42, 2016.

[5] T. Zhang, Y. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Cache-enabling UAV communications: Network deployment and resource allocation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7470–7483, 2020.

[6] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 188–195, 2018.

[7] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2017.

[8] X. Liu, J. Wang, N. Zhao, Y. Chen, S. Zhang, Z. Ding, and F. R. Yu, "Placement and power allocation for NOMA-UAV networks," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 965–968, 2019.

[9] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE network*, vol. 32, no. 2, pp. 24–31, 2018.

[10] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[11] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2018.

[12] P. D. Thanh, H. T. H. Giang, and I. Koo, "UAV-assisted NOMA downlink communications based on content caching," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 786–791, IEEE, 2020.

[13] J. Luo, J. Song, F.-C. Zheng, L. Gao, and T. Wang, "User-centric UAV deployment and content placement in cache-enabled multi-UAV networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 5, pp. 5656–5660, 2022.

[14] M. Amjad, L. Musavian, and S. Aissa, "NOMA versus OMA in finite blocklength regime: Link-layer rate performance," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16253–16257, 2020.

[15] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, 2014.