Hybrid NOMA for Latency Minimization in Wireless Federated Learning for 6G Networks

Pillappan KAVITHA, Kamatchi KAVITHA

Dept. of ECE, Velammal College of Engineering and Technology, Viraganoor, Madurai, Tamilnadu, India

pka@vcet.ac.in, kkavi@vcet.ac.in

Submitted June 16, 2023 / Accepted September 21, 2023 / Online first November 13, 2023

Abstract. *Wireless Federated Learning (WFL) is an innovative machine learning paradigm enabling distributed devices to collaboratively learn without sharing raw data. WFL is particularly useful for mobile devices that generate massive amounts of data but have limited resources for training complex models. This paper highlights the significance of reducing delay for efficient WFL implementation through advanced multiple access protocols and joint optimization of communication and computing resources. We propose optimizing the WFL Compute-then-Transmit (CT) protocol using hybrid Non-Orthogonal Multiple Access (H-NOMA). To minimize and optimize latency for the transmission of local training data, we use the Successive Convex Optimization (SCA) method, which efficiently reduces the complexity of non-convex algorithms. Finally, the numerical results verify the effectiveness of H-NOMA in terms of delay reduction, compared to the benchmark that is based on Non-Orthogonal Multiple Acces (NOMA).*

Keywords

WFL, NOMA, SCA, latency, Compute-then-Transmit (CT)

1. Introduction

Federated learning (FL), introduced in 2017 [1], is a distributed learning approach that Google pioneered. Federated learning (FL) has attracted substantial attention from both the realms of research and practical application [2], [3]. Unlike conventional distributed machine learning techniques, FL's primary objective is to achieve a unified, shared machine learning model for all participating distributed clients, while safeguarding the confidentiality of each client's sensitive data. The uniqueness of FL's privacy-preserving characteristic lies in its utilization of secure model aggregation, as opposed to the more conventional data aggregation. In the FL framework, individual clients train their machine-learning models locally and transmit only their model parameters to a central server without revealing their local data.

Several research studies have explored FL and its potential applications. Yi Liu et al. [4] described an overview of integrating federated learning into 6G communications and the core challenges of federated learning for 6G applications. In [5], the authors outlined a learning paradigm at the edge within distributed networks and conducted a comparison with traditional distributed data center computing and classical privacy-preserving learning. To provide communication efficiency, i.e. reduce uplink communication cost, two updates such as structured updates and sketched updates for communicating the local model to the central server in a federated learning have been provided in [6].

In [7], the authors discussed open problems and challenges present in FL and also described how FL gained traction in interdisciplinary fields such as machine learning, optimization, information theory, and statistics to cryptography, fairness, and privacy. D. Chen et al. [8] focused on how to solve the computation efficiency, low-latency object detection, and classification problems in augmented reality applications. In reliable federated learning, to select trusted mobile devices i.e., to guard against unreliable model updates, reputation has been introduced as a reliable metric [9]. Data privacy leakage issues related to ensuring secure FL in 5G networks have been addressed [10] and also proposed a blockchain-based framework to defend against poisoning attacks.

For collaborative model training at mobile edge networks, how FL can serve as an enabling technology has been presented in [11], and the authors also discussed the implementation of FL for privacy-preserving mobile edge networks. In [12], an overview of the integration of FL and blockchain, known as FLchain, in mobile edge networks is presented. Furthermore, the paper explored the utilization of FLchain for various applications, including edge data sharing, edge content caching, and edge crowdsensing. In [13], the authors explored insights into the implementation of distributed learning over wireless networks.

It is envisioned that 6G will heavily depend on pervasive artificial intelligence services and progressively surpass the capabilities of the fifth generation (5G) of wireless networks. Machine Learning (ML) has optimized wireless network

DOI: 10.13164/re.2023.0594

performance and enhanced data-driven applications [14]. Centralized configurations of ML techniques, which typically rely on a single entity like a central server for data upload and processing, could be more practical for upcoming 6G-enabled applications such as smart grids, autonomous vehicles, and augmented reality. This is due to strict latency requirements and concerns about data privacy. Consequently, the convergence of these constraints alongside the growing computational capabilities of devices has opened doors to the adoption of distributed frameworks for building learning models.

Although many works have successfully discussed the challenges imposed by FL, the authors in [15] provided an extensive analysis of the wireless aspect within the framework of federated learning (FL), concurrently enlightening the prospective paths for future research in the domain of WFL, aligned with the 6G vision. They also highlighted that WFL's efficacy is closely related to the capabilities of the underlying wireless communication network. Furthermore, their work emphasized that in the model transmission phase of each communication round, all devices employ a multiple access scheme to upload their individually trained outcomes to facilitate the seamless integration of WFL within the context of 6G.

In recent years, Non-Orthogonal Multiple Access (NOMA) has gained significant attention as a spectral-efficient multiple access technique [16]. In addition to its spectral efficiency benefits, NOMA has the potential to increase the number of served devices and provide fairness among users. This makes NOMA-based schemes a promising alternative for next-generation multiple access schemes, which are essential to meet the connectivity requirements of 6G [17]. In [18], the authors investigated the examination of a novel power allocation strategy designed to amplify the total throughput within a downlink NOMA system. In [19], the authors focused on the combination of NOMA and Mobile Edge Computing (MEC), and they also considered the application of NOMA uplink transmission to MEC, which enables multiple users to perform offloading simultaneously.

To minimize delay for offloading in a multi-user MEC network under maximum power, and energy constraints, the authors focused on NOMA and they also demonstrated that using NOMA can achieve a lower delay than Time Division Multiple Access (TDMA) under maximum power constraints [20]. In [21], the authors considered the minimization of the offloading delay for Nonorthogonal Multiple Access Assisted Mobile Edge Computing (NOMA-MEC). To minimize the energy consumption of the network, NOMA has been integrated with MEC networks in an underlay Unmanned Aerial Vehicle (UAV) [22]. For a wireless-powered MEC, the authors investigated the application of User Cooperation (UC) and NOMA [23]. In [24], the authors demonstrated that NOMA has the potential to decrease latency during a WFL round and accelerate the training process, which is essential for efficiently integrating WFL into 6G.

Our work explores the available potential of hybrid NOMA/OMA configurations to enhance the scalability of WFL. The notion of employing such hybrid NOMA configurations has already been introduced as a promising tactic for optimizing the offloading of data in Mobile Edge Computing (MEC), as demonstrated in references [25] and [26]. Building upon this foundation, our work seeks to extend and adapt the concept of hybrid NOMA to tackle the unique scalability challenges posed by WFL.

In the framework of optimizing the Compute-Then-Transmit (CT) protocol with the hybrid Non-Orthogonal Multiple Access (NOMA) technique, the challenge of formulating a multi-objective optimization problem aimed at minimizing latency requirements is encountered. This optimization problem is non-convex in nature which poses a computational problem. Low-complexity successive optimization algorithm is introduced to address this computational challenge.

The paper is organized as follows. Section 2 describes the system model. Section 3 analyses the delay minimization of Hybrid NOMA. Section 4 discusses its numerical result. The paper concludes in Sec. 5.

2. System Model

Consider a WFL with M number of users and a Server/Base Station (BS) as shown in Fig. 1. Each user is indexed as m, where $m \in M = \{1, 2, \ldots, M\}$. Each user m has a local dataset $\mathfrak{D}_m$, where $\mathrm{D}_m = |\mathfrak{D}_m|$ are the total data samples.

During each i^{th} communication round, the process follows these steps until global model convergence is achieved:

- BS broadcasts the global parameter w_i to all users participating in the current round.
- Upon receiving the global model parameter, each user $m \in M$ trains their respective local model using their dataset. Subsequently, the user uploads the trained local parameter to the server.
- Once all local parameters have been received, the server aggregates them to update the global model parameter to w_{i+1}.

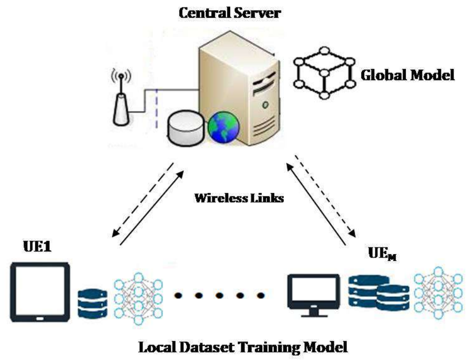

Fig. 1. System diagram.

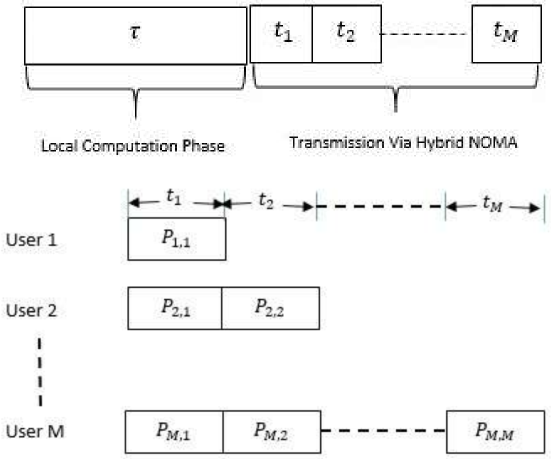

Fig. 2. Hybrid NOMA transmission phase.

As shown in Fig. 2, in Compute-Then-Transmit hybrid NOMA, two phases have been carried out during WFL round. In the first phase, users execute the local computations. While in the second phase, they transmit their messages, i.e., the trained parameters to the base station. A hybrid NOMA protocol has been considered during the transmission phase. In the hybrid NOMA strategy, some of the users are capable to complete their transmission before the other users, while the BS decodes by utilizing Successive Interference Cancellation (SIC). In Compute-Then-Transmit hybrid NOMA, all users are required to complete their local computations before the information transmission phase starts.

Unlike NOMA, in hybrid NOMA users close to the base station are allowed to complete their transmission before the other users. During the first t_1 seconds, user 1 is requested to complete its transmission. In addition to user 1, the other users are also permitted to carry out transmission during the first t_1 seconds. During the next t_2 seconds, user 2 is requested to complete its transmission, where the other users, user m, $2 \leq m \leq M$, can continue their transmission simultaneously. During the last t_M seconds, only user M is transmitted, because all other users should have already completed their transmission by then. Denote the user's transmit powers during t_n by $P_{m,1}$, $1 \leq n \leq m, 2 \leq m \leq M$.

By applying the proposed hybrid NOMA scheme, at t_n, the base station receives the following data from all user's trained data:

$$y_n = \sum_{m=n}^{M} \sqrt{P_{m,n}} h_m d_m^{-\frac{\alpha}{2}} s_{m,n} + \omega_n \tag{1}$$

where $s_{m,n}$ denotes the local trained data by user m during t_n, d_m is the distance from the user m to BS, α is the path loss exponent, the complex random variable $h_m \sim \mathcal{CN}(0,1)$ is the small scale fading and ω_n denotes the white Gaussian noise. The Successive Interference Cancellation (SIC) decoding is considered. The decoding process starts with the user closest to the base station and proceeds to decode the next user who is closer in proximity. At t_n, the base station first decodes user m's signal, when $m > n$, i.e., the signal from the user who is closer to the BS, before decoding signal, i.e., the signal from the user far away from the BS.

Using this SIC decoding order guarantees that user n experiences the same performance as with Orthogonal Multiple Access (OMA) at t_n. As a result, at t_n, user m's signal is decoded in the $(M-m+1)-th$ SIC step with the following data rate:

$$R_{m,n} = \log_2\left(1 + \frac{P_{m,n}|h_m|^2 d_m^{-\alpha}}{\sum_{j=n}^{m-1} P_{j,n}|h_j|^2 d_j^{-\alpha} + N_0 B}\right). \tag{2}$$

As a consequence, at t_n, user n's signal is decoded last with the data rate of $R_{n,n} = \log_2\left(1 + \frac{P_{n,n}|h_n|^2 d_n^{-\alpha}}{N_0 B}\right)$, which means that user n experiences interference-free information as in OMA.

The utilized computation resources for local model training, i.e., the CPU cycle frequency, for the m-th user is denoted as f_m. Let c_m represent the number of CPU cycles required for the m-th user to perform one sample of data and D_m represent the total data samples in local model training. Hence, the computation time dedicated to a local iteration from [24] and [27] is given as

$$\tau_m = \frac{c_m D_m}{f_m}, \quad \forall m \in M. \tag{3}$$

Accordingly, the energy consumption for a local iteration can be expressed as follows.

$$E_m^{\text{comp}} = \zeta c_m D_m f_m^2, \quad \forall m \in M \tag{4}$$

where ζ is a constant parameter related to the hardware architecture of device m. As discussed previously, all users are forced to complete the local computations within τ, with the corresponding energy consumed by each user being a decreasing function concerning τ. Thus, it should hold

$$\tau_m = \tau, \quad \forall m \in M. \tag{5}$$

3. Delay Minimization

The primary goal of this work is to minimize the overall delay experienced by all users during a WFL round by optimizing parameter aggregation on the server, improving processing capabilities, enhancing network communication, and implementing efficient algorithms. In the CT-Hybrid NOMA protocol, users terminate the computation phase simultaneously, but users complete its transmission phase at different times. So, the total delay of a WFL round is described as

$$T = \tau + t_1 + t_2 + \cdots + t_M. \tag{6}$$

The total delay of a WFL round is the sum of computation and transmission latency. This includes the time required for both the computation processes and the transmission of data. It is important to note that the delay of the server in broadcasting the global parameter is ignored, as the transmit power of the Base Station (BS) is significantly higher than that of the individual users. The BS transmits the same message to all users simultaneously without significant delay. Thus, the focus is primarily on minimizing the computation and transmission latency to reduce the overall delay experienced by the users. Also, we assume that the data size transmitted by each user is the same i.e., $Z_m = Z$ for $\forall m \in M$, related to the model parameters.

The CPU clock speed of each user is restricted to a maximum of $f_m^{\max}$. So, it must hold $f_m \leq f_m^{\max}, \forall m \in M$, which is equivalent to $\tau \geq \max\limits_{m \in M} \left(\frac{c_m D_m}{f_m^{\max}} \right) \triangleq b_1$. Moreover, the maximum available energy of each user is $E_m^{\max}$. So, it must hold

$$E_m^{\text{comp}} + E_m = \zeta \frac{c_m^3 D_m^3}{\tau^2} + E_m \leq E_m^{\max}, \quad \forall m \in M, \tag{7}$$

since the total consumed energy for both computation and communication purposes cannot exceed the maximum available energy.

The optimization problem for minimizing the latency of a WFL round in the case of CT-hybrid NOMA can be given as,

$$\begin{aligned} \min_{\tau, t_1, t_2, \ldots, t_M, E} \quad & \tau + t_1 + t_2 + \cdots + t_M, \\ \text{s.t.} \quad & C1 : \zeta \frac{c_m^3 D_m^3}{\tau^2} + E_m \leq E_m^{\max}, \forall m \in M, \\ & C2 : Z \leq \sum_{n=1}^{m} t_n R_{n,m}, \\ & C3 : \sum_{n=1}^{m} t_n P_{n,m} = E_m \end{aligned} \tag{8}$$

where $C1$, $C2$ and $C3$ are contraints. This optimization problem is a multi-objective optimization problem and the problem is non-convex because the objective and constraint functions are coupled with t_m and $P_{n,m}$. This paper proposes a low-complexity successive convex algorithm to solve this problem.

From (8), energy of the m^{th} user E_m can be written as

$$E_m = E_m^{\max} - \zeta \frac{c_m^3 D_m^3}{\tau^2}, \quad \forall m \in M. \tag{9}$$

Furthermore, since $E_m \geq 0, \forall m \in M$, by manipulating (9), it yields $\tau \geq \max\limits_{m \in M} \left(\sqrt{\zeta \frac{c_m^3 D_m^3}{E_m^{\max}}} \right) \triangleq b_2$. Thus, by also recalling that $\tau \geq b_1$, it should finally hold for τ

$$\tau \geq \max\{b_1, b_2\} \triangleq \tau_{\text{low}}. \tag{10}$$

3.1 Low-Complexity Successive Convex Optimization Algorithm

In this section, first, a low-complexity successive algorithm is proposed to solve the problem in (8). The Successive Interference Cancellation (SIC) algorithm is proposed to decode the data of users $1, 2, \ldots, M$. Due to the use of SIC, user n's choices for t_n and $P_{n,m}$ have no impact on user m's data rate, $m < n$. An extreme example is user 1's data rate, which is $R_{1,1} = t_1 B \log_2 \left[1 + \frac{P_{1,1} |h_1|^2 d_1^{-\alpha}}{N_0 B} \right]$ and depends only on t_1 and $P_{1,1}$.This motivates the use of a successive optimization strategy, where user m's transmission parameters are optimized after user $(m-1)$'s.

After manipulating τ from (10), $t_1, t_2, \ldots, t_M$ for minimizing overall delay can be solved by first solving for t_1. For user 1 maximum data size at the time, t_1 can be written as

$$Z = t_1 B \log_2 \left[1 + \frac{P_{1,1} |h_1|^2 d_1^{-\alpha}}{N_0 B} \right] \tag{11}$$

where $P_{1,1} = \frac{E_{1,1}}{t_1} = \frac{E_1}{t_1}$ and $A_1 = E_1 |h_1|^2 d_1^{-\alpha}$. So, the equation (11) becomes

$$Z = t_1 B \log_2 \left[1 + \frac{A_1}{t_1 N_0 B} \right]. \tag{12}$$

The optimal value of t_1 can be written from Appendix A by solving (12) as

$$t_1^* = - \frac{Z \ln(2) A_1}{B \left(Z N_0 \ln(2) + \mathcal{W}_{-1}(a_1) A_1 \right)} \tag{13}$$

where $\mathcal{W}_{-1}(\cdot)$ denotes Lambert W function and a_1 is given by

$$a_1 = - \frac{Z N_0 \ln 2}{A_1} 2^{-\frac{Z N_0}{A_1}}. \tag{14}$$

After performing global optimization, the ideal value of the user 1 delay is determined by optimizing the parameter τ. With the preparation completed, we can now proceed to implement the bisection method within the designated interval and present Algorithm 1 for acquiring the most favorable solutions that minimize the communication round delay considering only user 1.

Algorithm 1. Delay t_1 minimization for CT-Hybrid NOMA.

```
1: Initialize $\tau_{\text{low}} = \max\{b_1, b_2\}$, $\tau_m$, $\bar{\tau}$, $\tau_{\text{up}} = T_1(\bar{\tau})$, $\epsilon$;
2: while $\tau_{\text{up}} - \tau_{\text{low}} > \epsilon$ do
3:     Set $\tau = \tau_m$, and derive $t_1^*(\tau_m)$ from (13)
4:     Set $T_1(\tau_m) = \tau_m + t_1^*(\tau_m)$;
5:     Set $\tau = \tau_{\text{up}}$, and derive $t_1^*(\tau_{\text{up}})$ from (13)
6:     Set $T_1(\tau_{\text{up}}) = \tau_m + t_1^*(\tau_{\text{up}})$;
7:     if $T_1(\tau_m) < T_1(\tau_{\text{up}})$ then
8:         $\tau_{\text{up}} = \tau_m$
9:     else
10:        $\tau_{\text{low}} = \tau_m$
11:    end if
12:    $\tau_m = \frac{\tau_{\text{low}} + \tau_{\text{up}}}{2}$
13: end while
14: Output $\tau^* = \tau_{\text{up}}, t_1^* = t_1^*(\tau_{\text{up}}), T_1^* = \tau^* + t_1^*$
```

Subsequently, the algorithm's key steps are explained, starting with the required initializations in line 1, followed by the application of the bisection method in lines 2–13. In the subsequent lines 3–6 of the algorithm, the delay of the communication round considering user 1 is computed at the values of $\tau = \tau_m$ and $\tau = \tau_{\text{up}}$. Subsequently, in lines 7–12, the algorithm adjusts the bounds of τ appropriately in each iteration by comparing the values of $T_1(\tau_m)$ and $T_1(\tau_{\text{up}})$, aiming for convergence to the optimal solution.

To optimize $t_2, t_3, \ldots, t_M$, the optimization in (8) can be decomposed into subproblems, and defined as follows,

$$\begin{aligned} &\min_{t_m} t_m \\ \text{s.t.}\quad & C1: \zeta \frac{c_m^3 D_m^3}{\tau^2} + E_m \leq E_m^{\max}, 1 < m < M, \\ & C2: Z \leq \sum_{n=2}^{m} t_n R_{n,m}, \\ & C3: \sum_{n=2}^{m} t_n P_{n,m} = E_m. \end{aligned} \tag{15}$$

Algorithm 2 demonstrates how the proposed low-complexity successive convex algorithm tackles the multi-objective optimization problem in (8) by decomposing it into a sequence of subproblems described in (15) and solving them successively.

Algorithm 2. Low-complexity optimization algorithm.

```
Set $t_1^*$ from (13) and $P_{1,1}^* = \frac{E_1}{t_1^*}$;
$m = 1$.
while $m < M$ do
    $m = m + 1$.
    find the optimal solutions for $t_m^*, P_{m,n}, 1 \leq n \leq M$,
    by solving the problem
end while
The outcome of the algorithm is given by
$x^* \triangleq [t_2^*, t_3^*, \ldots, t_M^*, P_2^{*\text{T}}, P_3^{*\text{T}}, \ldots, P_M^{*\text{T}}]$
where $P_m^* = [P_{m,1}^*, P_{m,2}^* \ldots P_{m,m}^*]^{\text{T}}$.
```

Algorithm 2 can be interpreted as a greedy approach that breaks down problems in (8) into a set of subproblems, represented as (15), and sequentially solves them. The benefit of Algorithm 2 lies in its ability to achieve low computational complexity. While the solution obtained through Algorithm 2 is anticipated to be suboptimal for the problem in (8).

To this end, the major complexity of Algorithm 1 lies in applying the bisection method to derive t_1^* from (13). As a result, the complexity can be expressed as the order of $O\left(\log_2\left(\frac{\tau_{\text{up}} - \tau_{\text{low}}}{\epsilon}\right)\right)$. BS only needs to carry out $M - 1$ steps to implement Algorithm 2, where each step is to find the optimal solution to problems described in (15), and the associated computational complexity is moderate.

4. Numerical Results and Discussion

In this section, the performance of the proposed CT-Hybrid NOMA scheme is studied. This evaluation is conducted using MATLAB simulations, as detailed in Appendix B. The performance of CT-Hybrid NOMA is assessed based on the average latency achieved, which allows for a comparison with CT-NOMA in terms of information-theoretic perspectives while considering fading statistics. The CT-NOMA-based protocol [24] serves as the benchmark against which we assess and contrast the CT-NOMA protocol's efficacy in diminishing delays. To evaluate the performance of CT-Hybrid NOMA, simulation settings are taken from [24] as summarized in Tab. 1.

Figures 3 and 4 illustrate the influence of the maximum available energy of users on the average latency during a WFL round.

Parameter	Value
CPU cycle frequency ($f_m^{\max}$)	1.5 GHz
Bandwidth (B)	1.2 MHz
Path loss coefficient (α)	3.5
Parameter related to the hardware architecture (ζ)	10^{-27}
Number of users (M)	10 users
Total data samples of m^{th} user (D_m)	0.5 Mbit
Power spectral density (N_0)	–174 dBm/Hz
Number of CPU cycles required for the m^{th} user (c_m)	$\sim \mathcal{U}(10, 40)$
Distance between m^{th} user and BS (d_m)	$\sim \mathcal{U}(0, 1000\,\text{m})$

Tab. 1. Simulation settings.

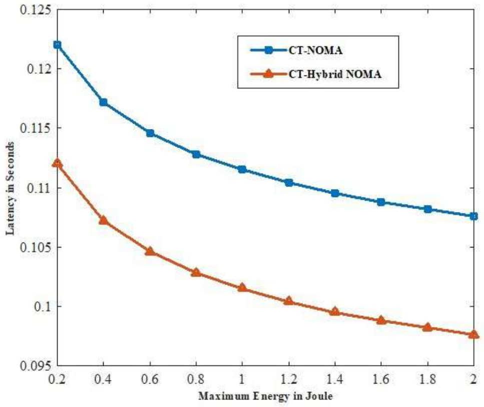

Fig. 3. Impact of the user's maximum available energy on latency, with $Z = 0.3$ Mbits.

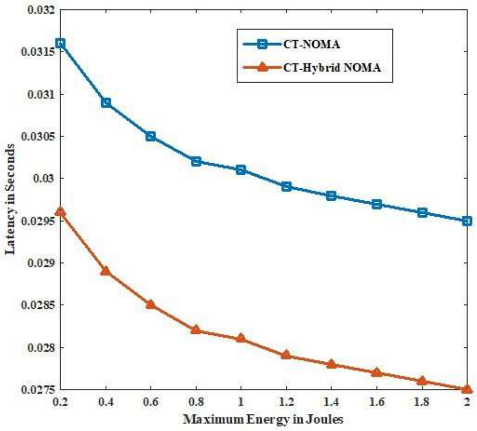

Fig. 4. Impact of the user's maximum available energy on latency, with $Z = 0.2$ Mbits.

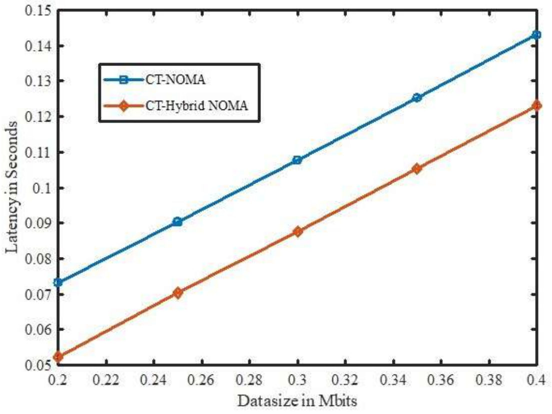

Fig. 5. Impact of the user's parameter data size on latency, with maximum energy of the user = 2 joule.

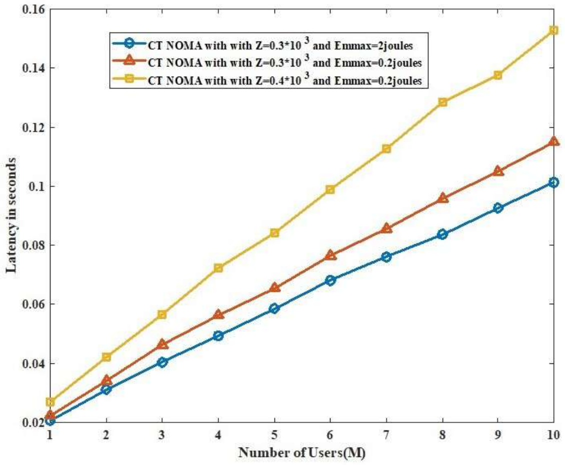

Fig. 6. Impact of the number of users on latency for CT NOMA protocol.

Analysis of Figs. 3 and 4 reveals the superior performance of CT-Hybrid NOMA over CT-NOMA. It's worth noting that CT-Hybrid NOMA achieves significantly better energy efficiency than CT-NOMA while maintaining comparable latency levels. In other words, CT-NOMA necessitates more energy to attain the same latency level as CT-Hybrid NOMA. The latency evaluation for both multiple access strategies encompasses two distinct data sizes, illustrated in Figs. 3 and 4 as 0.2 Mbits and 0.3 Mbits, respectively. Notably, an increase in data size to 0.3 Mbits corresponds to higher latency.

Figure 5 shows the influence of the different values of the data size of users on the average latency during a WFL round and it shows the superiority of CT-Hybrid NOMA over CT-NOMA across various users' data sizes. Figures 3 and 4 demonstrate a latency reduction of approximately 20 ms in CT-Hybrid NOMA when compared to CT-NOMA. Hybrid NOMA shows superior performance compared to the NOMA scheme by effectively utilizing the advantages of both orthogonal and non-orthogonal aspects of multiple access.

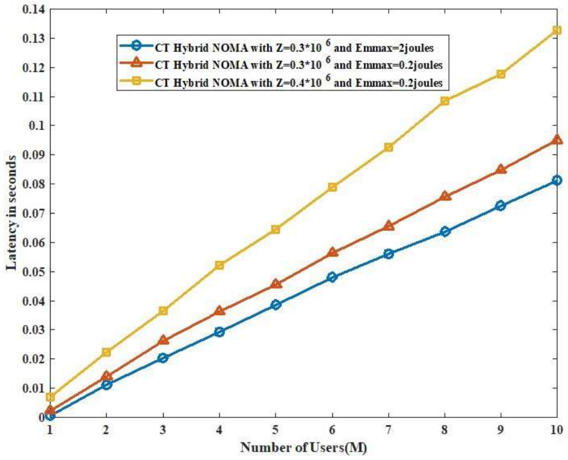

Fig. 7. Impact of the number of users on latency for CT Hybrid NOMA protocol.

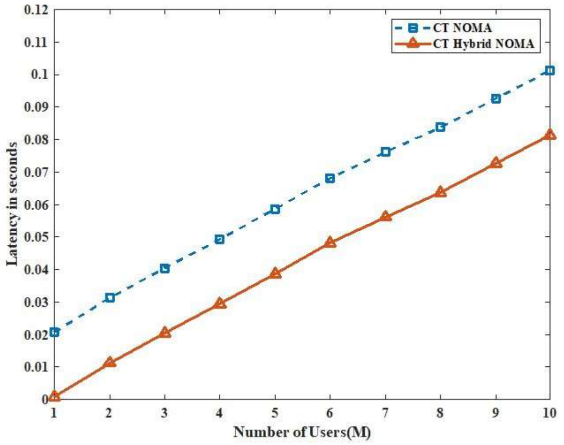

Fig. 8. Impact of the number of users on latency with $Z = 0.3 \times 10^6$ and $E_m^{\max} = 2$ joules.

Figures 6–8 show the impact of the number of users on latency. As the number of users increased, we observed a noticeable increase in latency for both the CT NOMA-based protocol and the CT Hybrid NOMA-based protocol. Interestingly, our results indicate that the CT Hybrid NOMA-based protocol consistently outperforms the CT NOMA-based protocol under all conditions. This suggests that the hybrid approach may offer better performance and lower latency as the user grows.

5. Conclusion

In this paper, a novel hybrid NOMA has been proposed to optimize the WFL Compute-then -Transmit(CT) protocol, accompanied by the formulation of a multi-objective optimization problem. Multi-objective optimization problem has been formulated to minimize the total communication round trip delay of users in WFL. Furthermore, in this paper, the SIC decoding order is solely based on the user's distance from the BS. Low-complexity SCA has been proposed to solve multi-objective optimization problems. Despite the high complexity involved in solving the optimization problem for implementing Hybrid NOMA in the Compute-then-Transmit (CT) protocol to optimize WFL, it is apparent that CT-Hybrid NOMA can reduce latency compared to CT-NOMA. This reduction in latency can significantly speed up the training process, which is a crucial necessity for effectively integrating WFL in 6G.

The low-complexity SCA method has been introduced to resolve non-convex optimization challenges within WFL. Deep Reinforcement Learning (DRL) is a promising path to further advance optimization. By utilizing DRL, resource management and decision-making in WFL systems have access to a powerful toolkit that can effectively solve complex problems. The goal of this convergence is to provide solutions to complex problems that are both nearly optimal and computationally feasible. This approach allows for practical and manageable resolutions to be achieved.

References

[1] MCMAHAN, B. H., MOORE, E., RAMAGE, D., et al. Communication-efficient learning of deep networks from decentralized data. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ft. Lauderdale (FL, USA), 2017, p. 1273–1282. DOI: 10.48550/arXiv:1602.05629

[2] CHEN, M., YANG, Z., SAAD, W., et al. A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 2021, vol. 20, no. 1, p. 269–283. DOI: 10.1109/TWC.2020.3024629

[3] YANG, Z., CHEN, M., WONG, K. K., et al. Federated learning for 6G: Applications, challenges, and opportunities. *Engineering*, 2022, vol. 8, p. 33–41. DOI: 10.1016/j.eng.2021.12.002

[4] LIU, Y., YUAN, X., XIONG, Z., et al. Federated learning for 6G communications: Challenges, methods, and future directions. *China Communications*, 2020, vol. 17, no. 9, p. 105–118. DOI: 10.23919/JCC.2020.09.009

[5] LI, T., SAHU, K., TALWALKAR, A., et al. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 2020, vol. 37, no. 3, p. 50–60. DOI: 10.1109/MSP.2020.2975749

[6] KONECNY, J., MCMAHAN, H. B., YU, F. X., et al. Federated learning: Strategies for improving communication efficiency. *arXiv*, 2016, p. 1–10. DOI: 10.48550/arXiv.1610.05492

[7] KAIROUZ, P., MCMAHAN, H. B., AVENT, B., et al. *Advances and Open Problems in Federated Learning*. Now Publishers, 2021. ISBN: 9781680837889

[8] CHEN, D., XIE, L. J., KIM, B., et al. Federated learning-based mobile edge computing for augmented reality applications. In *Proceedings of International Conference on Computing, Networking and Communications (ICNC)*. Big Island (HI, USA), 2020, p. 767–773. DOI: 10.1109/ICNC47757.2020.9049708

[9] KANG, J., XIONG, Z., NIYATO, D., et al. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 2020, vol. 27, no. 2, p. 72–80. DOI: 10.1109/MWC.001.1900119

[10] LIU, Y., PENG, J., KANG, J., et al. A secure federated learning framework for 5G networks. *IEEE Wireless Communications*, 2020, vol. 27, no. 4, p. 24–31. DOI: 10.1109/MWC.01.1900525

[11] LIM, W. Y. B., LUONG, N. C., HOANG, D. T., et al. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys and Tutorials*, 2020, vol. 22, no. 3, p. 2031–2063. DOI: 10.1109/COMST.2020.2986024

[12] NGUYEN, D. C., DING, M., PHAM, Q., et al. Federated learning meets blockchain in edge computing: Opportunities and challenges. *IEEE Internet of Things Journal*, 2021, vol. 8, no. 16, p. 12806–12825. DOI: 10.1109/JIOT.2021.3072611

[13] CHEN, M., GUNDUZ, D., HUANG, K., et al. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 2021, vol. 39, no. 12, p. 3579–3605. DOI: 10.1109/JSAC.2021.3118346

[14] LETAIEF, K. B., CHEN, W., SHI, Y., et al. The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine*, 2019, vol. 57, no. 8, p. 84–90. DOI: 10.1109/MCOM.2019.1900271

[15] BOUZINIS, P. S., DIAMANTOULAKIS, P. D., KARAGIANNIDIS, G. K. Wireless federated learning (WFL) for 6G networks part I: Research challenges and future trends. *IEEE Communications Letters*, 2022, vol. 26, no. 1, p. 3–7. DOI: 10.1109/LCOMM.2021.3121071

[16] DING, Z., LEI, X., KARAGIANNIDIS, G. K., et al. A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE Journal on Selected Areas in Communications*, 2017, vol. 35, no. 10, p. 2181–2195. DOI: 10.1109/JSAC.2017.2725519

[17] LIU, Y., ZHANG, S., MU, X., et al. Evolution of NOMA toward next generation multiple access (NGMA) for 6G. *IEEE Journal on Selected Areas in Communications*, 2022, vol. 40, no. 4, p. 1037–1071. DOI: 10.1109/JSAC.2022.3145234

[18] SREENU, S., KALPANA, N. Innovative power allocation strategy for NOMA systems by employing the modified ABC algorithm. *Radioengineering*, 2022, vol. 31, no. 3, p. 312–322. DOI: 10.13164/re.2022.0312

[19] DING, Z., FAN, P., POOR, H. V. Impact of non-orthogonal multiple access on the offloading of mobile edge computing. *IEEE Transactions on Communications*, 2019, vol. 67, no. 1, p. 375–390. DOI: 10.1109/TCOMM.2018.2870894

[20] ZENG, M., NGUYEN, N. P., DOBRE, O. A., et al. Delay minimization for NOMA-assisted MEC under power and energy constraints. *IEEE Wireless Communications Letters*, 2019, vol. 8, no. 6, p. 1657–1661. DOI: 10.1109/LWC.2019.2934453

[21] DING, Z., NG, D. W. K., SCHOBER, R., et al. Delay minimization for NOMA-MEC offloading. *IEEE Signal Processing Letters*, 2018, vol. 25, no. 12, p. 1875–1879. DOI: 10.1109/LSP.2018.2876019

[22] BUDHIRAJA, I., KUMAR, N., TYAGI, S., et al. Energy consumption minimization scheme for NOMA-based mobile edge computation networks underlying UAV. *IEEE Systems Journal*, 2021, vol. 15, no. 4, p. 5724–5733. DOI: 10.1109/JSYST.2021.3076782

[23] SU, B., NI, Q., YU, W., et al. Optimizing computation efficiency for NOMA-assisted mobile edge computing with user cooperation. *IEEE Transactions on Green Communications and Networking*, 2021, vol. 5, no. 2, p. 858–867. DOI: 10.1109/TGCN.2021.3056770

[24] BOUZINIS, P. S., DIAMANTOULAKIS, P. D., KARAGIANNIDIS, G. K. Wireless federated learning (WFL) for 6G networks part II: The compute-then-transmit NOMA paradigm. *IEEE Communications Letters*, 2022, vol. 26, no. 1, p. 8–12. DOI: 10.1109/LCOMM.2021.3121067

[25] DING, Z., XU, D., SCHOBER, R., et al. Hybrid NOMA offloading in multi-user MEC networks. *IEEE Transactions on Wireless Communications*, 2022, vol. 21, no. 7, p. 5377–5391. DOI: 10.1109/TWC.2021.3139932

[26] DURSUN, Y., FANG, F., DING, Z. Hybrid NOMA based MIMO offloading for mobile edge computing in 6G networks. *China Communications*, 2022, vol. 19, no. 10, p. 12–20. DOI: 10.23919/JCC.2022.00.024

[27] DURSUN, Y., GOKTAS, M. B., DING, Z. Green NOMA based MU-MIMO transmission for MEC in 6G Networks. *Computer Networks*, 2023, vol. 228, p. 1–7. DOI: 10.1016/j.comnet.2023.109749

About the Authors . . .

Pillappan KAVITHA was born in India, in 1979. She received the B.E. Degree in Electronics and Communication Engineering from Bharathidasan University, India, in 2002, and the M.E in Communication Systems from Thiagarajar College of Engineering, Madurai, India in 2009 and Ph.D. in Information and Communication from Anna University, Chennai in 2019. She has 15 and half years of teaching experience. Her current research interests include Wireless networks, MIMO OFDM, Probability and Stochastic analysis etc. She is a Life Member of the Indian Society for Technical Education (ISTE).

Kamatchi KAVITHA was born in India, in 1982. She received the B.E. Degree in Electronics and Communication Engineering from the Madurai Kamaraj University, India, in 2003, M.E in Wireless Technologies from Thiagarajar College of Engineering, Madurai, India in 2007 and Ph.D. in Information and Communication from Anna University, Chennai in 2016. She has 16 years of teaching experience. Her current research interests include MIMO OFDM, FSO, UWOC, etc. She is a Life Member of the Indian Society for Technical Education (ISTE).

Appendix A: Derivation for the Optimum Value of t_1 in (13)

From (12),

$$Z = t_1 B \log_2\left[1+\frac{A_1}{t_1 N_0 B}\right]. \tag{A1}$$

$y\mathrm{e}^y = w$ can be solved for y. Now, $y = \mathcal{W}_{-1}(w)$ where $\mathcal{W}_{.}$ is Lambert W function (source https://en.wikipedia.org/wiki/Lambert_W_function).

(A1) can be written as

$$Z = \frac{t_1 B}{\ln 2}\log_{\mathrm{e}}\left[1+\frac{A_1}{t_1 N_0 B}\right]. \tag{A2}$$

From (A2),

$$1+\frac{A_1}{t_1 N_0 B} = \mathrm{e}^{\frac{Z\ln 2}{t_1 B}}. \tag{A3}$$

Assume $1+\frac{A_1}{t_1 N_0 B} = x$, after some manipulations, we got

$$-\frac{xZN_0\ln 2}{A_1}\mathrm{e}^{-\frac{xZN_0\ln 2}{A_1}} = -\frac{ZN_0\ln 2}{A_1}2^{-\frac{ZN_0}{A_1}}. \tag{A4}$$

Consider the right hand side of (A4) as a_1 i.e., $-\frac{ZN_0\ln 2}{A_1}2^{-\frac{ZN_0}{A_1}} = a_1$ and using Lambert W function the solution to the above equation is given as

$$-\frac{xZN_0\ln 2}{A_1} = \mathcal{W}_{-1}(a_1). \tag{A5}$$

The optimal value of t_1 can be obtained from above as

$$t_1^* = -\frac{Z\ln(2)A_1}{B\left(ZN_0\ln(2)+\mathcal{W}_{-1}(a_1)A_1\right)}. \tag{A6}$$

Appendix B: MATLAB Code for Mathematical Model

Listing 1. Code for algorithm.

```
% CPU cycle frequency for the mth user
f_max=1.5*10^9;
B=1.2*10^6; % bandwidth
a=3.5; %path loss coefficient
% parameter related to the hardware architecture
c_tow=10^-27;
% Number of users
N=10;
% Data size of mth user
Dn=0.5*10^6;
  %the power spectral density (−174dBm)
No=-174;
% conversion from dBm to magnitude
No=10^(No/10)/1000;
%The no. of cycles for the n−th user
%local model training
cn=randi([10,40],1,N);
% diatance of nth user
dn=randi([0,1000],1,N);
% Channel coefficient
hn=1/sqrt(2)*(randn(1,N)+j*randn(1,N));
hn=sort(abs(hn));
nn=1;
Z=0.3*10^6; % Data Size
En_max=2;
[dn,index_d]=sort(dn);
% Block of codes to find tow_low value
for n=1:N
    c1(n)=cn(n)*Dn/f_max;
    c2(n)=sqrt(c_tow*(cn(n)^3)*(Dn^3)/En_max);
    g(n)=(hn(n)^2)*(dn(n))^-a;
end
a1=max(c1);
a2=max(c2);
tow_low=max(a1,a2);
% Block of lines to find t1 following algorithm 1
tow_dilt=0.05;
tow_m=tow_dilt;
[g,index]=sort(g,'descend');
%calling sub function t_tow_dilt
t_tow_dilt=abs(calculate_t1(g(1),cn(index(1)),En_max,
tow_m,Z,c_tow,Dn,No,B));
tow_up=tow_dilt+t_tow_dilt;
e=0.002;
ss=1;
while tow_up-tow_low>e
```

```
    tow=tow_m;
    %calling sub function t_tow_m
    t_tow_m=abs(calculate_t1(g(1),cn(index(1)),En_max,
   tow,Z, c_tow,Dn,No,B));
    tow=tow_up;
   %calling sub function t_tow_up
    t_tow_up=abs(calculate_t1(g(1),cn(index(1)),En_max,
   tow,Z,c_tow,Dn,No,B));
    T_m=tow_m+t_tow_m;
    T_up=tow_up+t_tow_up;
    if T_m<T_up
        tow_up=tow_m;
    else
        tow_low=tow_m;
    end
    tow_m=(tow_up+tow_low)/2;
end
tow_optimal=tow_up;
% calling sub function for calculating optimal t1 value
t1=abs(calculate_t1(g(1),cn(index(1)),En_max,
tow_optimal,Z,c_tow,Dn,No,B));

% sub function to calculate t1
function t1=calculate_t1(g1,cn1,En_max,tow,Z,c_tow,Dn,No,B)
EE1=(En_max-(c_tow*(cn1^3)*(Dn^3)/(tow^2)));
zn=Z;
num=-zn*log(2)*EE1*g1;
bn=-(2^(-zn*No/(EE1*g1)))*zn*No*log(2)/(EE1*g1);
W=lambertw(-1,bn);
den=B*((zn*No*log(2))+(W*EE1*g1));
t1=num/den;
% Energy of users not exceeding maximum value
for i=1:N
    EE(i)=En_max-(c_tow*(cn(index(i))^3)*(Dn^3)/(tow_optimal^2));
end
p=[];
p(1,1)=EE(1)/t1;
p(2,1)=p(1,1)*(g(1)^2)/(g(2)^2);
p(2,2)=((p(1,1)*((g(1))^2))+(p(2,1)*((g(2)^2))))/((g(2))^2);
t2_low=0.00001;
t2_up=t1-0.1*t1;
t2=calculate_t2(t1,t2_low,t2_up,p,EE,Z,g,No,B);
% Sub function to calculate t2 time
function t2=calculate_t2(t1,t2_low,t2_up,p,EE,Z,g,No,B)
R21=B*log2(1+(p(2,1)*g(2)/(p(1,1)*g(1)+(No*B))));
R22=B*log2(1+(p(2,2)*g(2)/(No*B)));
C1=t1*R21+t2_low*R22;
C2=t1*p(2,1)+t2_low*p(2,2);
while C1<=Z&&C2>=EE(2)&&t2_low<t2_up
    t2_low=t2_low+((t2_up-t2_low)*rand(1));
    R21=B*log2(1+(p(2,1)*g(2)/(p(1,1)*g(1)+(No*B))));
    R22=B*log2(1+(p(2,2)*g(2)/(No*B)));
    C1=t1*R21+t2_low*R22;
    C2=t1*p(2,1)+t2_low*p(2,2);
end
t2=t2_low;
```