Wireless Federated Learning (WFL) for 6G Networks - Part II: The Compute-then-Transmit NOMA Paradigm

Pavlos S. Bouzinis, *Student Member, IEEE*, Panagiotis D. Diamantoulakis, *Senior Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—As it has been discussed in the first part of this work, the utilization of advanced multiple access protocols and the joint optimization of the communication and computing resources can facilitate the reduction of delay for wireless federated learning (WFL), which is of paramount importance for the efficient integration of WFL in the sixth generation (6G) of wireless networks. To this end, in this second part we introduce and optimize a novel communication protocol for WFL networks, that is based on non-orthogonal multiple access (NOMA). More specifically, the *Compute-then-Transmit NOMA (CT-NOMA)* protocol is introduced, where users terminate concurrently the local model training and then simultaneously transmit the trained parameters to the central server. Moreover, two different detection schemes for the mitigation of inter-user interference in NOMA are considered and evaluated, which correspond to fixed and variable decoding order during the successive interference cancellation process. Furthermore, the computation and communication resources are jointly optimized for both considered schemes, with the aim to minimize the total delay during a WFL communication round. Finally, the simulation results verify the effectiveness of CT-NOMA in terms of delay reduction, compared to the considered benchmark that is based on time-division multiple access.**

***Index Terms*—Wireless Federated Learning, Non-Orthogonal Multiple Access (NOMA), Delay Minimization**

I. Introduction

In Part I of this two parts paper, we presented the advantages, the reference architecture and three core applications of wireless federated learning (WFL), the main challenges toward its efficient integration in the sixth generation (6G) of wireless networks, and the identified future directions [1]. This second part is motivated by the fact that, as it has been thoroughly discussed in II. A and III.A of the first part, the utilization of advanced multiple access protocols and the joint optimization of the communication and computing resources can facilitate WFL to meet the stringent latency requirements of 6G networks [1]–[3]. To this direction, the use of non-orthogonal multiple access (NOMA) has been proposed in the first part, due to offering low-latency and improved fairness by serving multiple users in the same resource block, in opposition to orthogonal multiple access (OMA), where each user occupies a single resource block [4].

The aim of the second part of this paper is to investigate the potential of using NOMA in WFL, motivated by the increased data rates that is capable of providing, which could possibly lead in reduced latency during the simultaneous transmission of the training parameters to the central server. In more detail, we introduce the *Compute-then-Transmit NOMA (CT-NOMA)* protocol, which is based on the use of two phases during a WFL round. Specifically, users firstly execute and complete the local model training concurrently and afterwards upload the trained model to the server simultaneously via NOMA. Moreover, we consider two detection schemes for the implementation of successive interference cancellation (SIC) process in NOMA that correspond to the utilization of a fixed or variable decoding order. The latter is implemented through the time-sharing (TS) strategy, which improves the performance of uplink NOMA by achieving any point of the multiple access channel (MAC) capacity region [5], and is the driving factor for accomplishing reduced delay. In this direction, we minimize the total delay of a WFL round subject to users' energy constraints for both considered schemes, by jointly optimizing the available computation and communication resources, i.e., users' central processing unit (CPU) frequency for local training, the transmission energy and the time intervals for local computations and parameter transmission. Finally, simulation results demonstrate the effectiveness of the proposed protocol in reducing the delay of a WFL round, compared with the considered benchmark that is based on time-division multiple access (TDMA).

P. S. Bouzinis, P. Diamantoulakis, and G.K. Karagiannidis are with Wireless Communication and Information Processing Group (WCIP), Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece, E-mails: {mpouzinis, padiaman, geokarag}@auth.gr

II. System Model and Proposed Protocols

A. Wireless federated learning model

Based on the reference architecture that has been presented in detail in the first part of this work, we consider a WFL learning system which consists of N users indexed as $n \in \mathcal{N} = \{1, 2, ..., N\}$ and a server/base station (BS). Each user n has a local dataset $\mathcal{D}_n$, where $D_n = |\mathcal{D}_n|$ are the total data samples. Next, we briefly describe the steps of an arbitrary (i-th) communication round, which are repeated until the global model converges.

i) The BS broadcasts the global parameter $\boldsymbol{w}^i$ to all users during the considered round.
ii) After receiving the global model parameter, each user $n \in \mathcal{N}$, trains the local model through its dataset, and then uploads the trained local parameter $\boldsymbol{w}_n^{i+1}$ to the server.
iii) After receiving all the local parameters, the server aggregates them, in order to update the global model parameter $\boldsymbol{w}^{i+1}$.

B. Compute-then-Transmit NOMA

As it has already been mentioned, CT-NOMA is based on the use of two consecutive phases during a WFL round, i.e., the computation and the communication phase. Specifically, 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 BS. During the information transmission phase, the NOMA protocol is considered. According to NOMA, users are capable of transmitting their messages simultaneously, while the BS decodes the users' messages by utilizing SIC. As implied by the term CT-NOMA, all users begin the information transmission procedure via NOMA at the same time instant. Therefore, all users are constrained to complete the local computations before the information transmission phase, i.e., within time duration τ, while the transmission phase duration is denoted as t. The CT-NOMA protocol is illustrated in Fig. 1.

The utilized computation resources for local model training, i.e., the CPU cycle frequency, from the n-th user is denoted as f_n. The number of CPU cycles for the n-th user to perform one sample of data in local model training is denoted by c_n. Hence, the computation time dedicated for a local iteration is given as

$$\tau_n = \frac{c_n D_n}{f_n}, \quad \forall n \in \mathcal{N}. \tag{1}$$

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

$$E_n^{\text{comp}} = \zeta c_n D_n f_n^2 = \zeta \frac{c_n^3 D_n^3}{\tau_n^2}, \quad \forall n \in \mathcal{N}, \tag{2}$$

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

$$\tau_n = \tau, \quad \forall n \in \mathcal{N}. \tag{3}$$

In the continue, the two proposed schemes regarding information transmission will be described, termed as NOMA with fixed decoding order (FDO) and NOMA with TS.

1) NOMA with FDO: Fixed decoding order refers to the standard uplink NOMA protocol with SIC, i.e., the use of an unchangeable decoding order throughout the whole transmission phase, which is widely used in the literature [4]. Let π_n be the the n-th user's decoding order position at the BS, while without loss of generality we consider that $\pi_1 < \pi_2 < ... < \pi_N$. According to NOMA, for decoding the first user's message, interference is created by the residual users, i.e., $n = 2, ..., N$, while on the second user's message, interference is created by the users indexed as $n = 3, ..., N$ and so on. Finally, the last user to be decoded, i.e., user N, experiences no interference. Following that, the data size Z_n of the $\boldsymbol{w}_n$ model parameter, that the n-th user transmits within t time duration, should satisfy the following condition

$$Z_n \leq tB \log_2 \left(1 + \frac{\frac{E_n}{t} g_n}{\sum_{\{i \in \mathcal{N} | \pi_i > \pi_n\}} \frac{E_i}{t} g_i + BN_0} \right), \forall n \in \mathcal{N}, \tag{4}$$

where B is the available bandwidth, N_0 denotes the power spectral density, while E_n is the n-th user's consumed energy for the aforementioned transmission, with $\frac{E_n}{t}$ being the corresponding transmit power. Moreover, $g_n = |h_n|^2 d_n^{-a}$ denotes the channel gain, where the complex random variable $h_n \sim \mathcal{CN}(0, 1)$ is the small scale fading, a is the path loss exponent and d_n is the distance between user n and the BS. Finally, we assume that the channel remains constant during a WFL round and can be perfectly estimated by the BS.

2) NOMA with TS: The basic principle of this scheme is that the users' decoding order can change throughout the transmission time duration, i.e., t, unlike to the NOMA with FDO where a fixed decoding order is considered. According to [5], the capacity region which can be achieved by uplink NOMA with TS, known as the multiple access channel (MAC) capacity region, is described as [5]

$$\sum_{i \in \mathcal{S}} Z_i \leq tB \log_2 \left(1 + \frac{\sum_{i \in \mathcal{S}} E_i g_i}{tBN_0} \right), \quad \forall \mathcal{S} \subseteq \mathcal{N}, \mathcal{S} \neq \emptyset, \tag{5}$$

where $\mathcal{S}$ is a non-empty subset of $\mathcal{N}$. Furthermore, it deserves to be mentioned that (5) also determines the performance of rate splitting multiple access (RSMA), which is an alternative scheme to achieve any point of the MAC capacity region [6]. It is noted that, through the thorough exploitation of the capacity region, NOMA with TS is expected to contribute in achieving low latency in WFL.

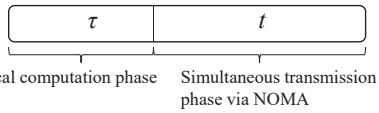

Fig. 1. *Compute-then-Transmit NOMA* during a WFL round.

III. Delay Minimization of a WFL Round

According to the WFL principles, all users experience the same delay, which is equal to the total delay of a WFL round, since the server needs all parameters in order to aggregate them. Thus, the main objective of this work, is to minimize the total delay of a WFL round. Since in the CT-NOMA protocol, users terminate both the computation and transmission phases simultaneously, the total delay of a WFL round is given as

$$T = \tau + t, \tag{6}$$

which is the summation of the computation and transmission latency. Note that, since the transmit power of the BS is much higher than that of the users' and the BS transmits the same message to all users, we ignore the delay of the server for broadcasting the global parameter. Hereinafter, we focus on a single arbitrary round, while the subsequent analysis can be similarly carried out for any WFL round. Finally, we assume that each user transmits the same data size $Z_n = Z$, $\forall n \in \mathcal{N}$, related with the model parameters.

The maximum available CPU clock speed of each user is denoted as $f_n^{\max}$. Thus, it should hold $f_n \leq f_n^{\max}$, $\forall n \in \mathcal{N}$, which is equivalent to $\tau \geq \max_{n \in \mathcal{N}} \left(\frac{c_n D_n}{f_n^{\max}} \right) \triangleq a_1$. Moreover, the maximum available energy of each user is $E_n^{\max}$. Hence, it should hold

$$E_n^{\text{comp}} + E_n = \zeta \frac{c_n^3 D_n^3}{\tau^2} + E_n \leq E_n^{\max}, \quad \forall n \in \mathcal{N}, \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2021.3121067, IEEE Communications Letters

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

A. CT-NOMA with time-sharing

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

$$\begin{aligned}\min_{\tau,\, t, \boldsymbol{E}} \quad & \tau + t \\ \textbf{s.t.} \quad & \mathrm{C}_1 : \zeta \frac{c_n^3 D_n^3}{\tau^2} + E_n \leq E_n^{\max}, \quad \forall n \in \mathcal{N}, \\ & \mathrm{C}_2 : \sum_{i \in \mathcal{S}} Z \leq tB \log_2 \left(1 + \frac{\sum_{i \in \mathcal{S}} E_i g_i}{tBN_0}\right), \quad \forall \mathcal{S} \subseteq \mathcal{N}, \\ & \mathrm{C}_3 : \tau \geq a_1,\, t \geq 0, E_n \geq 0,\, \forall n \in \mathcal{N},\end{aligned} \tag{8}$$

where C_2 is related to the users' achievable capacity region in (5). It is easy to verify that the problem in (8) is jointly convex with respect to (w.r.t.) $\tau, t, \boldsymbol{E}$ and thus can optimally be solved with standard convex-optimization solving methods, such as the interior-point. However, the computational cost to directly solve this by standard methods remains intractable, due to the large number of constraints in C_2. More specifically, the number of constraints in C_2 is equal to $2^N - 1$, since this is the total number of all non-empty subsets $\mathcal{S}$ of the set $\mathcal{N}$. As a result, solving (8) for large N comes at the expense of high complexity. To alleviate this burden, an efficient method for solving (8) will be proposed. Firstly, we consider a fixed value of τ. Since $\log(\cdot)$ is an ascending function w.r.t E_n, an optimal solution occurs when the constraints C_1 are satisfied with equality, i.e., users utilize their whole available energy, which leads to

$$E_n = E_n^{\max} - \zeta \frac{c_n^3 D_n^3}{\tau^2}, \quad \forall n \in \mathcal{N}. \tag{9}$$

Furthermore, since $E_n \geq 0, \forall n \in \mathcal{N}$, by manipulating (9), it yields $\tau \geq \max_{n \in \mathcal{N}} \left(\sqrt{\zeta \frac{c_n^3 D_n^3}{E_n^{\max}}}\right) \triangleq a_2$. Thus, by also recalling that $\tau \geq a_1$, it should finally hold for τ

$$\tau \geq \max\{a_1, a_2\} \triangleq \tau_{\text{low}}. \tag{10}$$

Next, after substituting (9) and considering a known $\tau \geq \tau_{\text{low}}$, the optimization problem in (8) can be reformulated as

$$\begin{aligned}\min_{t} \quad & t \\ \textbf{s.t.} \quad & \sum_{i \in \mathcal{S}} Z \leq tB \log_2 \left(1 + \frac{\sum_{i \in \mathcal{S}} A_i}{tBN_0}\right), \quad \forall \mathcal{S} \subseteq \mathcal{N},\end{aligned} \tag{11}$$

where A_i is given by

$$A_i = E_i g_i = \left(E_i^{\max} - \zeta \frac{c_i^3 D_i^3}{\tau^2}\right) g_i. \tag{12}$$

Note that A_i is constant for a given τ. Following that, without loss of generality we assume that $A_1 \geq A_2 \geq ... \geq A_N$. According to [7], when NOMA is combined with the TS scheme and by considering the aforementioned order of A_i, $\forall i \in \mathcal{N}$, the achievable capacity region of the users can be equivalently described as

$$\sum_{i=n}^{N} Z \leq tB \log_2 \left(1 + \frac{\sum_{i=n}^{N} A_i}{tBN_0}\right), \quad \forall n \in \mathcal{N}, \tag{13}$$

while by considering that $\sum_{i=n}^{N} Z = Z(N+1-n)$, (13) can be rewritten as

$$Z \leq \frac{tB \log_2 \left(1 + \frac{\sum_{i=n}^{N} A_i}{tBN_0}\right)}{(N+1-n)}, \quad \forall n \in \mathcal{N}. \tag{14}$$

It should be noted that the capacity region is now described by N constraints, i.e., Z is bounded by a set of N inequalities. Next, by exploiting the alternative, but still equivalent representation of the capacity region in (14), the optimization problem in (11) can be rewritten as

$$\begin{aligned}\min_{t} \quad & t \\ \textbf{s.t.} \quad & Z(N+1-n) \leq tB \log_2 \left(1 + \frac{\sum_{i=n}^{N} A_i}{tBN_0}\right), \forall n \in \mathcal{N}.\end{aligned} \tag{15}$$

Lemma 1: The optimal value of t for the problem in (15) can be written as

$$t^* = \max_{n \in \mathcal{N}} \left[-\frac{z_n \ln(2) \sum_{i=n}^{N} A_i}{B\left(z_n N_0 \ln(2) + \mathcal{W}_{-1}(b_n) \sum_{i=n}^{N} A_i\right)} \right], \tag{16}$$

where $z_n = (N+1-n)Z$ and b_n is given by

$$b_n = -\frac{2^{-\frac{z_n N_0}{\sum_{i=n}^{N} A_i}} z_n N_0 \ln(2)}{\sum_{i=n}^{N} A_i}, \tag{17}$$

while $\mathcal{W}_{-1}(\cdot)$ denotes the secondary branch of the Lambert W function.

Proof: We will show that the optimal t is given by the most stringent inequality, among the set of N inequalities in (15), when this hold with equality. Firstly, it is straightforward to show that the function

$$f(t) = tB \log_2 \left(1 + \frac{\sum_{i=n}^{N} A_i}{tBN_0}\right), \quad \forall t > 0, \tag{18}$$

is monotonically increasing with respect to t. Following that, we assume that t' is optimal and satisfies

$$Z(N+1-n) < t' B \log_2 \left(1 + \frac{\sum_{i=n}^{N} A_i}{t'BN_0}\right), \forall n \in \mathcal{N}, \tag{19}$$

i.e., all constraints are satisfied with strict inequality. It will be shown by contradiction that this is not an optimal solution for t. Since $f(t)$ is increasing w.r.t t, there exist an $t^* < t'$, for which at least one inequality constraint from (15) is satisfied with equality. This observation contradicts the fact that t' is optimal, since the objective is to minimize t. Therefore, t^* is given by the most stringent inequality, among the set of N inequalities in (15), when this hold with equality. After some mathematical manipulations, t^* can be written as in (16) and the proof is completed. ∎

As a matter of fact, the problem can be efficiently solved with closed form solutions for a fixed τ. In the continue, we will show that the global optimal solution can be obtained via the bisection method, in order to find the optimal value of τ. Firstly, it will be shown that τ^* is bounded, according to the following lemma.

Lemma 2: The optimal τ is bounded in the following interval

$$\tau_{\text{low}} \leq \tau^* \leq \tau_{\text{up}}, \tag{20}$$

where $\tau_{\text{up}} = T(\bar{\tau}) = \bar{\tau} + t^*(\bar{\tau})$ is the total delay of the communication round that corresponds to an arbitrary feasible value of τ, denoted by $\bar{\tau}$, while $t^*(\bar{\tau})$ denotes the optimal t for a fixed $\tau = \bar{\tau}$ and is given by (16).

Proof: As mentioned previously, the optimization problem is feasible when the condition $\tau \geq \tau_{\text{low}}$ is satisfied. Since $\bar{\tau}$ could be any feasible solution but not necessarily the optimal one, it holds $T(\bar{\tau}) \geq T(\tau^*)$. Following this, by considering that the dedicated time for computations cannot exceed the total delay for any feasible τ, i.e., $\tau < T(\tau)$, we conclude to $\tau^* < T(\tau^*) \leq T(\bar{\tau})$. Thus, the upper bound of τ^* can be expressed as $\tau_{\text{up}} = T(\bar{\tau})$ and the proof is completed. ■

As a result, we have derived the upper and lower bound of τ^* and now we are ready to apply the bisection method in the considered interval and propose Algorithm 1 for obtaining the optimal solutions that minimize the delay of a communication round. Next, the main steps of the proposed algorithm are discussed. After the necessary initializations in line 1, the bisection method is applied throughout lines 2-13. More specifically, in lines 3-6, the total delay of the communication round is calculated at the points $\tau = \tau_{\text{m}}$ and $\tau = \tau_{\text{up}}$. Following that, in lines 7-12, by comparing $T(\tau_{\text{m}})$ and $T(\tau_{\text{up}})$, the bounds of τ are properly adjusted in each iteration, until the convergence is achieved. It should be highlighted that the joint convexity of the primary problem in (8), w.r.t. all the considered variables, guarantees the convergence of the bisection method to the optimal solution. After the resolution of the optimization problem and the calculation of τ^*, t^*, the optimal E_n, f_n can be given as $E_n^* = E_n^{\max} - \zeta \frac{c_n^3 D_n^3}{\tau^{*2}}$ and $f_n^* = \frac{c_n D_n}{\tau^*}$. To this end, the major complexity of Algorithm 1 lies in applying the bisection method and in searching among N values, in order to derive t^* from (16). As a result, the complexity can be expressed as $\mathcal{O}\left(N \log_2\left(\frac{\tau_{\text{up}} - \tau_{\text{low}}}{\epsilon}\right)\right)$, where ϵ denotes the algorithm's tolerance.

Algorithm 1 Delay Minimization for CT-NOMA with TS

1: **Initialize** $\tau_{\text{low}} = \max\{a_1, a_2\}$, $\tau_{\text{m}} = \bar{\tau}$, $\tau_{\text{up}} = T(\bar{\tau})$, ϵ;
2: **while** $\tau_{\text{up}} - \tau_{\text{low}} > \epsilon$ **do**
3: **Set** $\tau = \tau_{\text{m}}$, and derive $t^*(\tau_{\text{m}})$ from (16)
4: **Set** $T(\tau_{\text{m}}) = \tau_{\text{m}} + t^*(\tau_{\text{m}})$;
5: **Set** $\tau = \tau_{\text{up}}$, and derive $t^*(\tau_{\text{up}})$ from (16)
6: **Set** $T(\tau_{\text{up}}) = \tau_{\text{up}} + t^*(\tau_{\text{up}})$;
7: **if** $T(\tau_{\text{m}}) < T(\tau_{\text{up}})$
8: $\tau_{\text{up}} = \tau_{\text{m}}$;
9: **else**
10: $\tau_{\text{low}} = \tau_{\text{m}}$;
11: **end if**
12: $\tau_{\text{m}} = \frac{\tau_{\text{low}} + \tau_{\text{up}}}{2}$;
13: **end while**
14: **Output** $\tau^* = \tau_{\text{up}}$, $t^* = t^*(\tau_{\text{up}})$, $T^* = \tau^* + t^*$;

B. CT-NOMA with fixed decoding order

The optimization problem for minimizing the total round delay in the case of CT-NOMA with FDO, can be written as

$$\begin{aligned}
\min_{\tau,\, t, \boldsymbol{E},\, \boldsymbol{\pi}} \quad & \tau + t \\
\textbf{\textit{s.t.}} \quad & \mathrm{C}_1 : \zeta \frac{c_n^3 D_n^3}{\tau^2} + E_n \leq E_n^{\max}, \quad \forall n \in \mathcal{N}, \\
& \mathrm{C}_2 : Z \leq tB \log_2 \left(1 + \frac{\frac{E_n}{t} g_n}{\sum_{\{i \in \mathcal{N} | \pi_i > \pi_n\}} \frac{E_i}{t} g_i + BN_0}\right) \\
& \quad \forall n \in \mathcal{N}, \\
& \mathrm{C}_3 : \tau \geq a_1,\, t \geq 0, E_n \geq 0,\, \forall n \in \mathcal{N},\, \boldsymbol{\pi} \in \Pi,
\end{aligned} \tag{21}$$

where the permutation $\boldsymbol{\pi}$ belongs to the set Π, defined as the set of all possible decoding orders among the N users. Note that the total number of permutations, i.e., all possible decoding orders, is equal to $|\Pi| = N!$. Therefore, the optimization problem in (21) is of a combinatorial nature. However, an exhaustive search among all possible decoding orders is prohibitive, due to high complexity. To this end, we adopt an ascending decoding order w.r.t. to the channel gains, according to which the messages of the users with weaker channel gains are exposed to less interference during the decoding process. It is notable that this a common selection for uplink NOMA systems, since it provides fairness among users without reducing the sum rate [7], [8]. Accordingly, by considering without loss of generality that $g_1 \geq g_2 \geq ... \geq g_N$, we set $\pi_1 < \pi_2 < ... < \pi_N$. As a result, C_2 of (21) can be reformulated as

$$Z \leq tB \log_2 \left(1 + \frac{E_n g_n}{\sum_{i=n+1}^{N} E_i g_i + tBN_0}\right), \quad \forall n \in \mathcal{N}. \tag{22}$$

By substituting (22) in (21), the later can be written as

$$\begin{aligned}
\min_{\tau,\, t, \boldsymbol{E}} \quad & \tau + t \\
\textbf{\textit{s.t.}} \quad & \mathrm{C}_1 : \zeta \frac{c_n^3 D_n^3}{\tau^2} + E_n \leq E_n^{\max}, \quad \forall n \in \mathcal{N}, \\
& \mathrm{C}_2 : Z \leq tB \log_2 \left(1 + \frac{E_n g_n}{\sum_{i=n+1}^{N} E_i g_i + tBN_0}\right), \\
& \quad \forall n \in \mathcal{N}, \\
& \mathrm{C}_3 : \tau \geq a_1,\, t \geq 0, E_n \geq 0,\, \forall n \in \mathcal{N},
\end{aligned} \tag{23}$$

which is non-convex due to the coupling of t and $\boldsymbol{E}$ in constraint C_2. In order to address the non-convexity issues, we first rewrite C_2 as

$$2^{\frac{Z}{tB}} \leq \frac{E_n g_n}{\sum_{i=n+1}^{N} E_i g_i + tBN_0}, \forall n \in \mathcal{N}. \tag{24}$$

In the continue we set $E_n \triangleq \exp(\tilde{E}_n), \forall n \in \mathcal{N}$ and $t \triangleq \exp(\tilde{t})$. After some mathematical manipulations, C_2 can be written as

$$\begin{aligned}
& \ln\left(2^{\frac{Z}{B} \exp(-\tilde{t})} - 1\right) \\
& + \ln\left(\sum_{i=n+1}^{N} \exp(\tilde{E}_i - \tilde{E}_n) g_i + \exp(\tilde{t} - \tilde{E}_n) BN_0\right) \\
& \leq \ln(g_n), \quad \forall n \in \mathcal{N}.
\end{aligned} \tag{25}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2021.3121067, IEEE Communications Letters

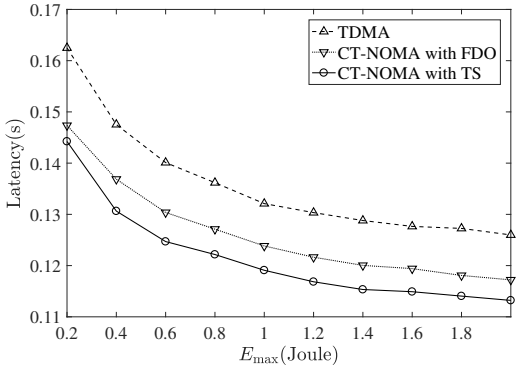

Fig. 2. Impact of the users' maximum available energy on latency, with $Z = 0.3$Mbits.

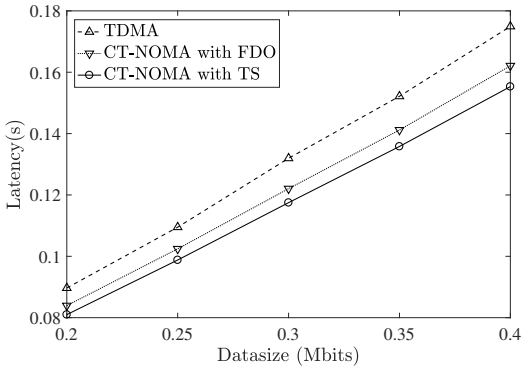

Fig. 3. Impact of the users' parameter data size on latency, with $E_n^{\max} = 2$ Joule.

By exploiting the aforementioned formulation of C_2, the optimization problem in (23) can be transformed to the following equivalent one:

$$\begin{aligned}
&\min_{\tau,\,\tilde{t},\tilde{\boldsymbol{E}}} \quad \tau + \exp(\tilde{t}) \\
&\textbf{\textit{s.t.}} \quad \mathrm{C}_1 : \zeta \frac{c_n^3 D_n^3}{\tau^2} + \exp(\tilde{E}_n) \leq E_n^{\max}, \quad \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C}_2 : \ln\left(2^{\frac{Z}{B}\exp(-\tilde{t})} - 1\right) + \ln\Bigg(\sum_{i=n+1}^{N} \exp(\tilde{E}_i - \tilde{E}_n) g_i \\
&\qquad\qquad + \exp(\tilde{t} - \tilde{E}_n) B N_0\Bigg) - \ln(g_n) \leq 0, \, \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C}_3 : \tau \geq a_1, \, \tilde{t} \in \mathbb{R}, \tilde{E}_n \in \mathbb{R}.
\end{aligned} \tag{26}$$

It can be easily shown that the optimization problem in (26) is convex, while the proof is omitted for brevity. Thus, it can efficiently be solved with standard convex optimization methods. We finally highlight that the optimization problems as well as the SIC in NOMA, are executed by the central server, which is equipped with powerful computational capabilities.

IV. Performance Evaluation and Discussion

In this section, CT-NOMA is evaluated in terms of the achieved average latency, which is indicative of the achieved performance when comparing different multiple access schemes from an information-theoretic perspective, taking into account the fading statistics. Thus, the following figures have been extracted by means of Monte Carlo, by using the simulations' parameters setup that is summarized in Table I.

In Fig. 2, the impact of the users' maximum available energy on the average latency during a WFL round is depicted. We use the TDMA-based protocol as benchmark and compare it with the CT-NOMA protocol in terms of delay reduction, taking into account both considered detections schemes for NOMA. It can been observed that CT-NOMA with FDO outperforms TDMA. Moreover, CT-NOMA with TS clearly dominates both TDMA and CT-NOMA with FDO. This is due to the fact that CT-NOMA with TS can achieve any point of the capacity region, in contrast to CT-NOMA with FDO which achieves only the corner points. As a result, the efficient exploitation of the capacity region, as well as subsequently the flexible interplay among users' data rates that CT-NOMA with TS enables, lead to decreased delay during the WFL round. Also, it deserves to be noted that CT-NOMA greatly outperforms TDMA in terms of energy consumption for the same latency.

TABLE I
Simulation Settings

Parameter	Value	Parameter	Value
$f_n^{\max}$	1.5 GHz	D_n	0.5 Mbit
B, a	1.2MHz, 3.5	N_0	-174 dBm/Hz
ζ	10^{-27}	c_n(cycles)	$\sim \mathcal{U}(10, 40)$
N	10 users	d_n	$\sim \mathcal{U}(0, 1000\mathrm{m})$

In Fig. 3, the superiority of CT-NOMA with TS against CT-NOMA with FDO, and TDMA is again corroborated for different values of users' data size. Furthermore, CT-NOMA with FDO presents an enhanced performance in comparison with TDMA.

By evaluating the demonstrated results, it is evident that CT-NOMA has the potential to provide decreased latency during a WFL round and subsequently accelerate the training process, which is an important requirement for the efficient integration of WFL in 6G. Moreover, since NOMA can achieve any point of the capacity region through the use of time-sharing (or rate splitting) technique, it can serve in future research as a performance upper bound in order to evaluate other multiple access schemes for WFL. Furthermore, the proposed protocol could serve as a baseline for more sophisticated network implementations, such as hybrid NOMA/OMA configurations, which could possibly enhance the scalability of WFL.

References

[1] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G Networks - Part I: Research challenges and future trends," *submitted to IEEE Commun. Lett.*, 2021.

[2] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[4] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel.Areas in Commun.*, vol. 35, no. 10, pp. 2181–2195, 2017.

[5] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[6] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, 1996.

[7] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422–8436, 2016.

[8] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, 2016.