An Overview of Enabling Federated Learning over Wireless Networks

Fotis Foukalas [†], Athanasios Tziouvaras [*], Theodoros A. Tsiftsis [**†]

[†] Computer Science and Telecommunications, University of Thessaly, Lamia, Greece
[*] Electrical and Computer Engineering, University of Thessaly, Volos, Greece
[**] School of Intelligent Systems Science & Engineering, Jinan University, Zhuhai, China
Email: [*] foukalas@ieee.org

***Abstract*—In this paper, we provide an overview of enabling federated learning (FL) techniques over wireless networks. More specifically, we present key techniques such as model compression, quantization and sparsification that increase the training accuracy of the distributed learning over the wireless medium. Next, the joint FL, resource allocation and scheduling approach is presented, which is identified in two types: a) both user and network assisted, and b) network assisted only. More specifically, the proposed FL-driven resource allocation and scheduling result in a joint optimization problem, where resource allocation and scheduling are jointly optimized. Finally, the simulation setup is described and the obtained simulation results are discussed, while several key enabling techniques are employed that further highlight the achievable performance of enabling FL over wireless networks in terms of training accuracy and loss.**

***Index Terms*—Federated learning, wireless networks, resource allocation, scheduling, simulation.**

I. INTRODUCTION

Federated learning (FL) over wireless networks is a quite new topic, which needs special attention since FL was not designed to work over wireless medium. The new design challenge is considered the efficient co-design of FL with wireless networks, where the objective is to optimize the training accuracy of the multiple users' data. Towards this end, there are several challenges already identified in [1] and [2]. For example, radio resources play a key role on deploying FL over wireless networks efficiently. In particular, FL should be considered over wireless fading channels, where the low capacity of unreliable links need to be efficiently managed through quantization [3]. Other useful solutions are considered the accuracy-loss correction and model compression [4]. Further, quantization [5] and loss regularization techniques [6] aim to address the noisy wireless medium by specifying the loss function in a different way.

On the other hand, FL should take also into account resource allocation and scheduling policies [7-9]. In such a case, a joint solution must be provided given the fact that FL is evolved over the time [10]. For example, the latency determined by either considering learning accuracy or time and user's energy consumption is considered in case of FL over wireless networks [11]. In general, a joint learning, resource allocation and user selection optimization problem is formulated, where the goal is to minimize the FL loss function [12]. In such a joint optimization problem, the fast convergence of the overall procedure is also critical [13]. Further, multichannel random access is also considered for more efficient update uploading, which is the key factor for implementing FL over wireless transmissions [14].

It is evident from the literature above that enabling FL over wireless networks should address several design challenges in order to achieve high training accuracy and low loss. FL requirements could vary from low computation solutions, e.g. quantization, compression, and sparsification, to the design of resource allocation and scheduling policies depending on the application, e.g. smart city, intelligent transportation and immersive experience [2],[4]. In this work, we provide an overview and comparison of key techniques, which enable FL running over wireless networks efficiently. Such an overview has been not provided yet as made clear from the literature above, where in [1] and [7] the authors discussed several open issues without developing, simulating and comparing the most important key techniques as provided below in this paper. More specific, our contribution is considered to pledge the following research elements. We first present the FL modeling over wireless networks, which is used as the baseline to compare with the key techniques. Next, we focus on the key techniques such as the compression, quantization and sparsification, which are able to attain low communication overhead. Finally, we present a detailed co-design of the FL with resource allocation and scheduling. The simulation setup of the multi-user FL over a wireless network configuration is explained, which can be met to different wireless applications such as mobile users or Internet of Things (IoT) devices connected to a 5G network for example. Simulation results are finally presented, which highlight the achievable performance of the FL in case of efficient co-design with wireless networks.

The rest of this paper is organized as follows. Section II presents the FL model over wireless networks. Section III presents key techniques to enable FL over wireless networks. Section IV provides simulation setup and results, while Section V concludes this work.

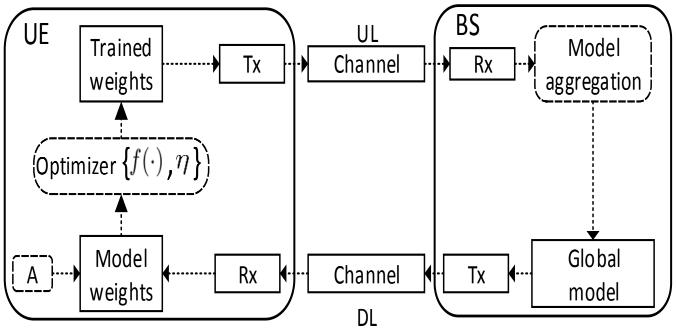


Fig. 1: Federated learning (FL) model over wireless networks.

II. Federated learning Modeling over Wireless Networks

A. Federated Learning Model

Fig. 1 depicts the model of FL over wireless networks. In particular, user equipments (UEs) train their local model independently. During this local training process, each model's weights and biases W are stimulated by the activation function (A) and updated by the optimizer as a result of the loss function $f(\cdot)$ minimization. A set of optimizers can be parameterized by setting the learning rate η parameter, which defines the speed by which the model adapts to the problem. The resulting trained model parameters W are transmitted from the UE to the base station (BS) through the transmitter (Tx) at the uplink (UL) channel [1]. In the sequel, the BS receives the W parameters through the Rx and aggregates them in one global model according to the following equation:

$$G(W) = \frac{1}{M}\sum_{i=0}^{M} w_i, \tag{1}$$

where M is the number of the UEs and w_i^{k+1} are the trained model parameters of the $i\text{-}th$ UE. Thus, the resulting global model $G(W)$ is an average of the UE parameters that participate in the training process. Finally, the BS sends back the global model $G(W)$ to the UEs through the Tx baseband processing so that the next federated round to begin [3].

B. Frame Transmission for Federated Learning

Due to bandwidth limitations, the amount of available resources is constrained and thus, each UE is not able to acquire the necessary resource blocks (RBs) for the FL process. The available RBs are allocated among the UEs that participate in the FL process according to the employed resource allocation scheme. To this end, each UE utilizes the allocated amount of RBs that are used to pass the trained model parameters or any other FL related information. Therefore, in case of FL over wireless networking, a UE is able to transmit an amount of information proportional to the amount of the allocated RBs. To this end, the model parameters are mapped into the available RBs and the radio frame is transmitted over-the-air to the BS. Table I presents the amount of RBs per bandwidth configuration and also the amount of frames required for the UEs to transmit the model parameters to the BS with 10 UEs. Limited bandwidth configurations may lead to resource competition and hence, the amount of UEs that may participate in each federated round will be limited too. To this end, scheduling policies are employed to coordinate the UE participation for each FL round. More specifically, such policies designate only a subset of UEs to transmit their local model updates to the BS each round while the rest of the UEs do not participate in the federated round. The arrangement and the amount of transmitting UEs changes each round according to the employed scheduling policy, as discussed below.

TABLE I: Frames per federated round.

Bandwidth	RB / Slot	RB / Frame	Frames / Federated round
1.4 MHz	6	120	8
3 MHz	15	300	7
5 MHz	25	500	6
10 MHz	50	1000	4
15 MHz	75	1500	2
20 MHz	100	2000	1

III. Enabling Federated Learning over Wireless Networks

A. Compression, Quantization and Sparsification

The limited available bandwidth per UE is a major factor that has a negative impact to the performance of the FL process [1], [12]. As a result, a number of sub-channels with limited capacity may lead to low accuracy and high convergence time of the FL global model [15]. In order to address such a problem, model quantization was devised, which compresses the size of model parameters by using low precision representations for their numerical entries [16]. Quantized models can be used for the transmission of either global [17] or local model parameters [18]. Further, sparsification [3] has been proposed, which reduces the FL communication cost by transmitting only a few parameters of a model [19], instead of broadcasting the whole model. Model architecture compression has also been considered [4], which removes some insignificant parameters of the model and thus, shrinking its size without affecting the overall accuracy [20].

Fig. 2 depicts the key techniques, which can be incorporated in wireless networks. In particular, model compression (C) is either applied directly to the model architecture and removes redundant neuron connections or replaces certain model weights with zeros in order to increase the sparsity of the weight matrix. Model compression techniques are employed for either local models (UE side) or global models (BS side) depending on the FL requirements. On the other hand, quantization (Q) is a technique that decreases the size of model weights without removing any model parameters. To achieve this, quantizers reduce the numerical precision of the weights by changing their binary representation. The quantized weights require less bits as

[1] The Tx is actually the baseband processing with the corresponding physical (PHY) layer implementation.

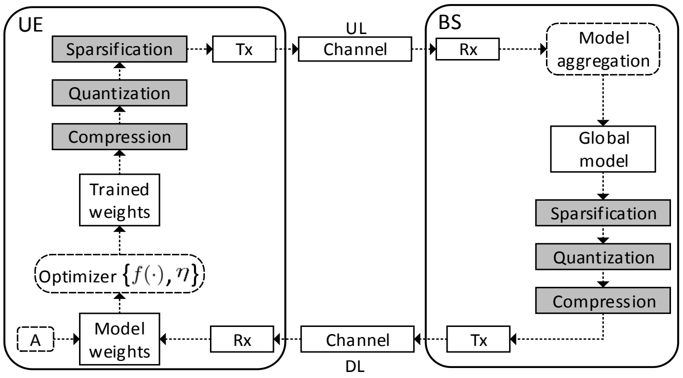


Fig. 2: Key techniques to enable FL over wireless networks.

a result of the information loss, which may also lead to loss of accuracy for small neural networks. Further, sparsification (S) is a very effective tool for reducing the size of the transmitted model parameters. By using sparsification, a UE or BS may omit to transmit certain model weights without decreasing the FL accuracy. The amount and type of model weights, which are excluded from the over-the-air transmission differs according to the employed sparsification technique.

B. Joint Resource Allocation and Scheduling

Due to radio resource limitations, efficient UE scheduling and resource allocation methods have been proposed in order to coordinate the global model updates for each FL round [9]. To this end, authors in [9] and [10] formed a loss optimization problem over UE scheduling, resource and power allocation subject to the signal-to-noise-ratio (SNR) of the UE. Further, in [3] and [16] authors formulated a loss optimization problem over local dataset samples of the UEs, UE scheduling and power allocation subject to the channel capacity and SNR of the UE. Such solutions utilize each UE separately in order to locally solve the corresponding optimization problem, while the BS is used for updating the necessary parameters and for broadcasting them to the UEs in order to calculate the solution. In contrast, other techniques employ the BS only for solving the joint resource allocation and scheduling optimization problem. More specific, in [13] and [21] the authors formed a loss optimization problem over RB allocation and scheduling and subject to the capacity and SNR of the uplink channel. In [8], authors also form a loss optimization problem over power allocation and subject to SNR, while in [11] and [22] researchers employed a loss optimization problem over RB and power allocation and subject to channel conditions such as SINR and channel capacity. Further, previous work in [23] authors optimized the loss function over the power and RB allocation with respect to the channel state information (CSI) reporting of the UEs. Given the solutions described above, resource allocation and scheduling can be classified into the two following categories:

1) **UE and Network assisted resource allocation:** UEs and BS take decisions for the resource allocation and scheduling in a cooperative way as in [9],[10]. We refer to such techniques as resource allocation and scheduling A (RAS-A).
2) **Network assisted resource allocation:** the BS manages the resource allocation and scheduling without the UEs participation as in [13], [15]. We refer to such techniques as resource allocation and scheduling B (RAS-B).

Fig. 3a depicts the RAS-A technique for efficient FL application. In particular, the UEs form and solve an optimization problem noted as P1 with regards to a set of constraints. Specifically, the P1 is formulated as a loss minimization problem over the RB allocation, scheduling and subject to power [5] allocation and learning rate [9], as follows :

$$P1 : argmin\{f(w_i^{k+1}) = \frac{1}{D_i}\sum_{i\in D} f(w_i^k), D_i, \eta_i^k\}, \qquad (2)$$

where $f(w_i^{k+1})$ is the loss function of the $i-th$ UE for the $k+1$ federated round, D_i the local data set of the $i\text{-}th$ UE and η_i^k the local optimizer learning rate. As a result, by minimizing the equation 2 in a federated way (i.e. with UE-BS cooperation) the loss function of the model is also minimized. Under this premise, each UE locally solves the P1 problem and dispatches the learning rate η_i^k along with the P1 solution and the trained model W to the BS. In the sequel, the BS aggregates the local parameters into a global model $G(W)$ and updates the learning rates of each UE according to the following formula [9]:

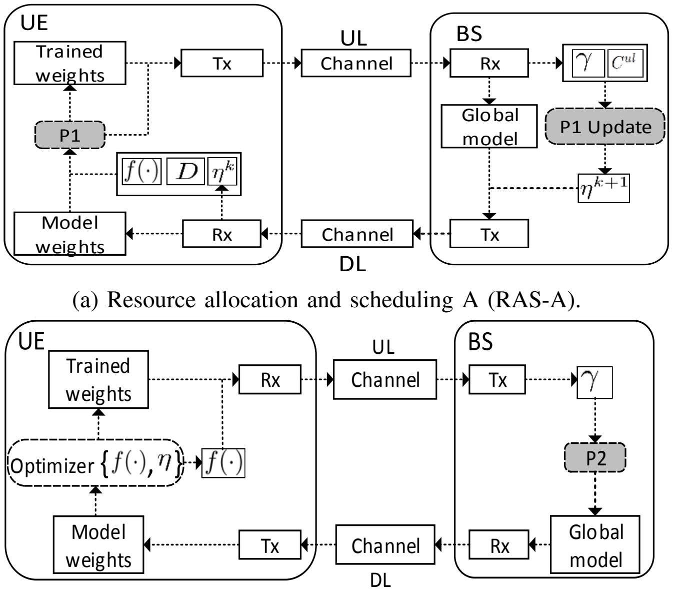
$$\eta_i^{k+1} = \frac{t*\eta_i^k}{t+1} + \frac{\sum_{m=1}^{M}\{1, \gamma_i^k > \theta\}}{t+1}, \qquad (3)$$

where k is the current federated round, θ is the SNR threshold under which the η_i^k is updated, M is the total number of UEs and γ_i^k is the SNR of the $i\text{-}th$ UE as in [10],[16]. The BS also calculates a vector a_i^{k+1} which is used for the UE *scheduling* process and corresponds to the following equation [3] [5]:

$$a_i^{k+1} = \frac{\gamma_i^{k+1}}{h_i^k(t)}, \qquad (4)$$

where $h_i^k(t)$is the channel gains vector as in [5] from the $i\text{-}th$ UE to the BS during the $k\text{-}th$ federated round. The resulting η_i^{k+1} and a_i^{k+1} values are broadcasted back to the corresponding UEs along with the aggregated global model $G(W)$. Each UE utilizes the new η_i values to minimize its optimization function for the next federated round while the new a_i vector indicates whether the i-th UE should participate on the next federated round. In this sense, the RB allocation and scheduling problem is solved locally at each UE while the BS assists this process by providing updates for the optimization parameters. As a result, UEs with better SNR tend to converge their local model faster compared to others due to the scaling of the learning rate η_i.

Fig. 3b depicts the RAS-B technique, in which the BS only manages the RB allocation of the UEs. In this case, the UE dispatch the local loss function [13] gradient and the trained model [22] to the BS. The BS in turn forms the following loss minimization problem, noted as P2, over the RB allocation and



(a) Resource allocation and scheduling A (RAS-A).

(b) Resource allocation and scheduling B (RAS-B).

Fig. 3: Joint FL, resource allocation and scheduling in wireless networks.

scheduling subject to the uplink capacity, SNR [8] and UE loss function:

$$P2 : argmin\{G^k(w) = \frac{1}{D_i M} \sum_{i \in M} \sum_{i \in D} f(w_i^k), \gamma_i^{k+1}, C^{ul}\}, \tag{5}$$

where M is the total amount of participating UEs, γ_i^k is the SNR of the $i\text{-}th$ UE ,C^{ul} the uplink channel capacity, $f(w_i^k$ the loss function of the $i\text{-}th$ UE. In Fig.3b, the BS solves the P2 minimization problem and proceeds in aggregating the local updates of the UEs which will have the highest impact to the global model loss reduction[21]. Next, it broadcasts the global model back to the UEs without transmitting any more parameters related with the optimization problem. Finally, the BS *schedules* and allocates RBs to the corresponding UEs according to their contribution to the eq.(5) minimization. As a result, the BS decides which UE to include on each federated round and thus it designates the amount of RBs that will be allocated to each device.

IV. Simulation Setup and Results

A. Simulation Setup

In order to implement the FL training process, we use the TensorFlow federated framework [24] in conjunction with the Matlab tool. We employ the TensorFlow for the deployment of the local training and the global aggregation techniques and Matlab for UE scheduling, resource allocation and simulation of the channel conditions under which the model parameters are transmitted. For the FL model, we employ the VGG-16 convolution deep neural network and the MNIST dataset which is composed of images of handwritten digits. The VGG-16 is a powerful DNN ideal for image classification and object recognition tasks [25] while the MNIST dataset is often used as a benchmark standard for the evaluation of machine learning techniques [26]. For the wireless network, we develop both the physical downlink shared channel (PDSCH) and the physical uplink shared channel (PUSCH) using Matlab.

We implement the compression (C), quantization (Q) and sparsification (S) FL techniques using the corresponding TensorFlow federated functions and the FL wireless networking techniques through Python functions. More specifically, for model compression, we employ a magnitude-based weight pruning technique which trims out insignificant model weights during the training process. For quantization, we opt to a Float16 post-training methodology, which reduces the bit size of the trained model weights to 16-bits. Regarding sparsification S, we employ a weight clustering approach which reduces the the number of unique weight values by grouping the weights of each layer into clusters. For the wireless network, we utilize both RAS-A and RAS-B techniques, which include the UE and network assisted resource allocation and schemes. The implementation is able to solve the two optimization problems, i.e. P1 and P2.

B. Simulation Results

Fig. 4 depicts the accuracy over federated rounds for two SNR values, i.e. $6dB$ and $7dB$, after employing quantization, sparsification, compression and regular (i.e. without C, Q or S) FL solutions [2]. We observe that the model compression C outperforms the rest of solutions and also achieves high accuracy (93%) under low SNR conditions. Further, quantization Q converges slower when compared to compression and but it high accuracy (92%) at the end of the training process when error-free transmission is considered. On the other hand, this technique is more prone to loss of accuracy due to lower SNR values, as its accuracy drops to 88% when SNR is $6dB$. Further, the sparsification technique appears to have the slowest convergence rate and the lowest accuracy (84.4%), when the SNR drops to $6dB$. Finally, the regular solution (RAS-R) performs very well under high SNR thresholds, as it does not employ a lossy compression technique and thus, the model parameters are transmitted uncompressed. We opt not to draw the RAS-R accuracy for the $6dB$ SNR as it drops at very low levels (44%) and cannot be properly illustrated in this figure.

Fig. 5 depicts the loss over federated rounds for different SNR values. We observe an identical performance as shown in 4, in reverse way though. The compression C technique achieves the best loss minimization (0.22), followed the quantization Q (0.23) and sparsification S (0.25) technique. A key difference is that the loss convergence is very slow for the sparsification S as the initial loss value is very high compared to the others. The compression C technique achieves the best performance as it successfully manages to eliminate insignificant weights that do not affect the model loss. Hence, it greatly reduces the amount of data transmitted over-the-air

[2] The regular FL is considered the one presented as a baseline FL model in Sec.II.

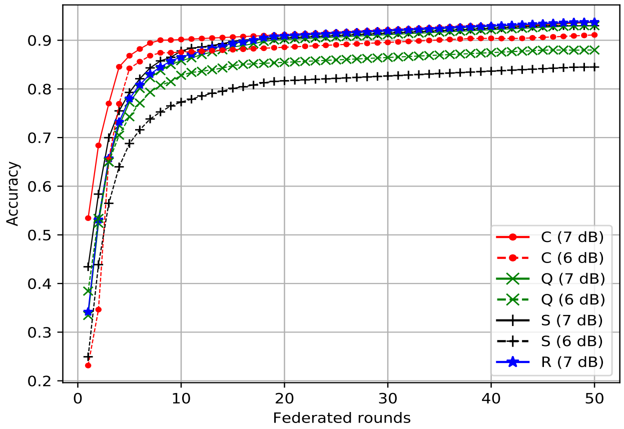


Fig. 4: Accuracy over federated round for different SNR values in dB and C, S, Q and R FL implementations.

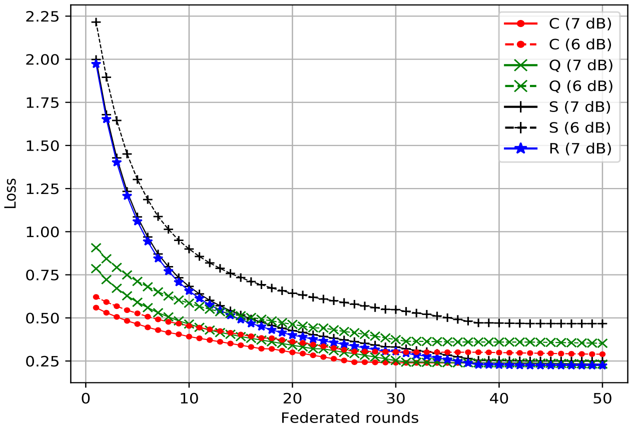


Fig. 5: Loss over federated round for different SNR values in dB and C, S, Q and R FL implementations.

and thus, the amount of transmission errors is kept at low levels. Similarly by employing the quantization Q technique, the communication costs between the UEs and the BS are greatly reduced as the quantized weights require significantly less bit for representation when compared with unquantized weights. The sparsification S technique also transmits lower amount of data over-the-air, but is more prone to bit-errors. This happens because the transmitted centroids contain values related with a great number of weights, and thus a small error on the transmitted centroid may affect the corresponding weight data. The regular methodology is not depicted in the picture under $6dB$ SNR, but it is outperformed by every other as it does not employ any techniques for error resilience or compression and thus, even a small bit error rate (BER) may significantly affect the loss function value. Specifically, RAS-R loss levels are 1.9 in the end of FL training process for $6dB$ SNR.

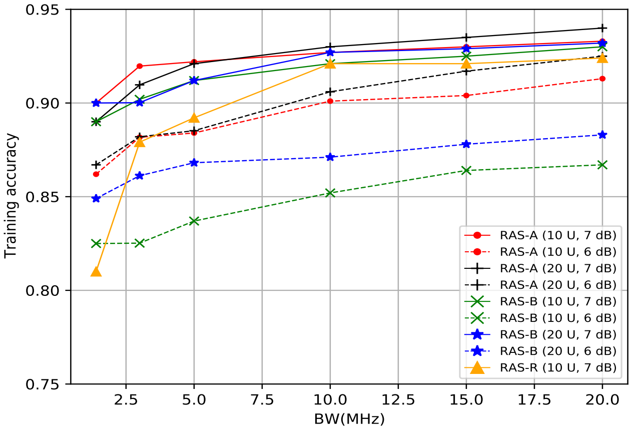


Fig. 6: Training accuracy over BW in MHz for RAS-A and RAS-B and different number of UEs and SNR in dB.

Fig. 6 depicts the accuracy over bandwidth for different UE number and SNR values. In this figure we compare the RAS-A and RAS-B techniques as described in section 3 that reflect on the methodologies followed by the existing literature. We observe that RAS-A performs better than RAS-B in general, as it manages to exploit the available network resources more efficiently while also depicting good BER resilience. RAS-A achieves 94% accuracy while RAS-B achieves 93% with 20 UEs and error-free channel transmission. Further, when the SNR drops to $6dB$ the RAS-A techniques manage to achieve 92% accuracy while the RAS-B solution accuracy drops to 88%. Also the amount of participating UEs plays an important role as more UEs result in less network resource allocation per UE. Under this premise, the RAS-A also outperforms in bandwidth limited environments even when the amount of UEs is high as they manage to coordinate efficiently the scheduling and RB allocation process. The regular type of RAS solution denoted as RAS-R performs better under high BW configurations, where each UE may be allocated the required RB for transmission but its accuracy drops very low when the RBs are limited. In Fig.6, the RAS-R achieves 36% accuracy in $1.4MHz$, 42% in $5MHz$ and 47% in $20MHz$, which we opt not to draw it in the figure for clarity reasons.

Fig. 7 plots the loss over bandwidth for different UE number and SNR values. We observe that a high amount of UEs and available bandwidth greatly contribute to the minimization of loss function. On the contrary, the combination of limited bandwidth environments with high amount of UEs lead to non optimal loss minimization due to the limited resource sharing of the UE. This problem is efficiently addressed by both RAS-A and RAS-B solutions which manage to achieve low loss values even under heavy resource constrains. We also observe that RAS-A solutions perform better in high bandwidth channels while RAS-B solutions perform very well in lower bandwidth configurations. It is shown that RAS-A and RAS-B achieve 0.27 and 0.25 correspondingly for 20 UEs and 1.4 MHz bandwidth while the loss values change to 0.22 and 0.25 in $20MHz$.

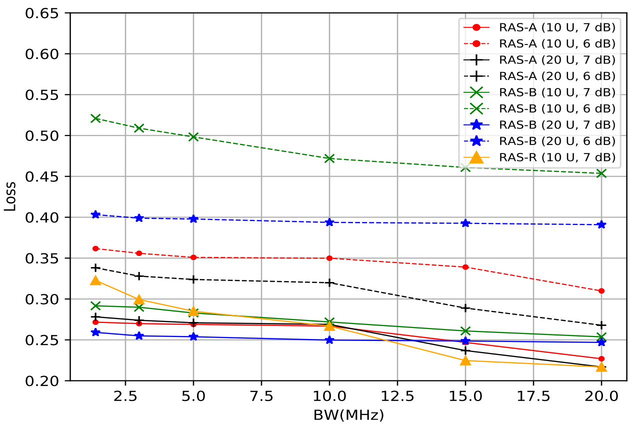


Fig. 7: Loss over BW in MHz for RAS-A and RAS-B and different number of UEs and SNR in dB.

The SNR levels also affect the RAS-B more than the RAS-A techniques due to the ability of RAS-A to adapt to worse channel conditions. Finally, the RAS-R technique performs very well under high bandwidth and SNR conditions but its performance drops with lower SNR or when the UE resources are limited. RAS-R achieves 1.9 loss for 20 UEs and $6dB$ under $1.4MHz$ bandwidth, 1.5 under $5MHz$ and 1.3 under $20MHz$.

V. Conclusion and Future Work

In this work, we presented an overview and comparison of enabling techniques for FL over wireless networks. To this end, we first presented a detailed model of FL over wireless networks. Next, we presented key techniques such as quantization, sparsification and compression mechanisms. Moreover, we designed a joint optimization in order to provide efficient resource allocation and scheduling assisted by both the UE and the network or the network only. A simulation setup is also explained and the obtained simulation results are demonstrated in order to highlight and compare the performance of each technique. In particular, the achievable performance is presented in terms of training accuracy and loss by using the key enabling techniques, which is compared with the regular FL model. Our future work would be the design of an end-to-end FL framework to enable high training accuracy for the FL-aware future radio access networks.

References

[1] S. Niknam, H. S. Dhillon and J. H. Reed, Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges, IEEE Commun. Magazine, vol. 58, no. 6, pp. 46-51, Jun. 2020.

[2] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang and M. Guizani, Reliable Federated Learning for Mobile Networks, IEEE Wirel. Commun., vol. 27, no. 2, pp. 72-80, Apr. 2020.

[3] M. M. Amiri and D. Gndz, Federated Learning Over Wireless Fading Channels, IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3546-3557, May 2020.

[4] Z. Zhao, C. Feng, H. H. Yang and X. Luo, Federated-Learning-Enabled Intelligent Fog Radio Access Networks: Fundamental Theory, Key Techniques, and Future Trends, IEEE Wireless Communications, vol. 27, no. 2, pp. 22-28, Apr. 2020.

[5] M. M. Amiri, D. Gunduz, S. R. Kulkarni, H. V. Poor, Convergence of Federated Learning over a Noisy Downlink, https://arxiv.org/abs/2008.11141, Aug. 2020.

[6] F. Ang, L. Chen, N. Zhao, Y. Chen and W. Wang and F. Richard Yu, Robust Federated Learning With Noisy Communication, IEEE Trans. on Communications, vol. 68, no. 6, pp. 3452-3464, Jun. 2020.

[7] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, K. Huang, Toward an Intelligent Edge: Wireless Communication Meets Machine Learning, IEEE Communications Magazine, vol. 58, no. 1, pp. 19-25, Jan. 2020.

[8] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor and S. Cui, "A Joint Learning and Communications Framework for Federated Learning over Wireless Networks", in arXiv, June 2019.

[9] H. H. Yang, Z. Liu, T. Q. S. Quek and H. V. Poor, "Scheduling Policies for Federated Learning in Wireless Networks," in IEEE Transactions on Communications, vol. 68, no. 1, pp. 317-333, Jan. 2020.

[10] M. M. Wadu, S. Samarakoon and M. Bennis, Federated Learning under Channel Uncertainty: Joint Client Scheduling and Resource Allocation, 2020 IEEE Wireless Communications and Networking Conference (WCNC), May 2020.

[11] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen and C. S. Hong, Federated Learning over Wireless Networks: Optimization Model Design and Analysis, IEEE INFOCOM - IEEE Conference on Computer Communications, Paris, France, pp. 1387-1395, May 2019

[12] M. Chen, Z. Yang, W. Saad, Ch. Yin, H. V. Poor and Sh. Cui, Performance Optimization of Federated Learning over Wireless Networks, 2019 IEEE Global Communications Conference (GLOBECOM), Dec. 2019.

[13] H. T. Nguyen, V. Sehwag, S. Hosseinalipour, C. G. Brinton, M. Chiang and H. V. Poor, Fast-Convergent Federated Learning, https://arxiv.org/abs/2007.13137, Jul. 2020.

[14] J. Choi and S. R. Pokhrel, Federated Learning With Multichannel ALOHA, IEEE Wirel. Communi. Letters, vol. 9, no. 4, pp. 499-502, Apr. 2020.

[15] R. Balakrishnan, M. Akdeniz, S. Dhakal and N. Himayat, Resource Management and Fairness for Federated Learning over Wireless Edge Networks, 2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).

[16] M. M. Amiri and D. Gunduz, Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air, 2019 IEEE International Symposium on Information Theory (ISIT), Jul. 2019.

[17] M. M. Amiri, D. Gunduz, S. R. Kulkarni and H. V. Poor, Federated Learning With Quantized Global Model Updates, https://arxiv.org/abs/2006.10672, Jun. 2020.

[18] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, S. Cui, Federated Learning with Quantization Constraints, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020.

[19] K. Yang, T, Jiang, Y. Shi and Z. Ding, Federated Learning via Over-the-Air Computation, IEEE Transactions on Wireless Communications, vol. 19, no. 3, pp. 2022-2035, Mar. 2020.

[20] Y. He, X. Zhang and J. Sun, "Channel Pruning for Accelerating Very Deep Neural Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 1398-1406, Dec. 2017

[21] M. Chen, H. V. Poor, W. Saad and Sh. Cui, Convergence Time Minimization of Federated Learning over Wireless Networks, IEEE International Conference on Communications (ICC), Jun. 2020.

[22] C. Dinh, N. H. Tran, M. N. H. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya and V. Gramoli, Federated Learning over Wireless Networks: Convergence Analysis and Resource Allocation, https://arxiv.org/abs/1910.13067, Mar. 2020.

[23] W. Shi, Sh. Zhou and Z. Niu, Device Scheduling with Fast Convergence for Wireless Federated Learning, ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Jun. 2020.

[24] TensorFlow federated framework, WebLink: https://www.tensorflow.org/federated

[25] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, June 2016.

[26] G. Patrini, A. Rozza, A. K. Menon, R. Nock and L. Qu, "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2233-2241, July 2017.