Scheduling Policies for Federated Learning in Wireless Networks: An Overview

SHI Wenqi, SUN Yuxuan, HUANG Xiufeng, ZHOU Sheng, NIU Zhisheng
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Due to the increasing need for massive data analysis and machine learning model training at the network edge, as well as the rising concerns about data privacy, a new distributed training framework called federated learning (FL) has emerged and attracted much attention from both academia and industry. In FL, participating devices iteratively update the local models based on their own data and contribute to the global training by uploading model updates until the training converges. Therefore, the computation capabilities of mobile devices can be utilized and the data privacy can be preserved. However, deploying FL in resource-constrained wireless networks encounters several challenges, including the limited energy of mobile devices, weak onboard computing capability, and scarce wireless bandwidth. To address these challenges, recent solutions have been proposed to maximize the convergence rate or minimize the energy consumption under heterogeneous constraints. In this overview, we first introduce the backgrounds and fundamentals of FL. Then, the key challenges in deploying FL in wireless networks are discussed, and several existing solutions are reviewed. Finally, we highlight the open issues and future research directions in FL scheduling.

Keywords: federated learning; wireless network; edge computing; scheduling

DOI: 10.12142/ZTECOM.202002003

https://kns.cnki.net/kcms/detail/34.1294.TN.20200610.1010.004.html, published online June 10, 2020

Manuscript received: 2020-02-10

Citation (IEEE Format): W. Q. Shi, Y. X. Sun, X. F. Huang, et al., "Scheduling policies for federated learning in wireless networks: an overview," *ZTE Communications*, vol. 18, no. 2, pp. 11 – 19, Jun. 2020. doi: 10.12142/ZTECOM.202002003.

1 Introduction

With the deployment of deep learning algorithms on Internet-of-Things (IoT) devices at the network edge[1] and the explosive growth of mobile data[2], technologies like edge learning[3] emerge and focus on running deep learning algorithms at the wireless access network. To ensure the performance of deep learning in practical scenarios, such as auto-driving and user preference prediction, efficient training of the learning model with the data generated at the network edge is necessary. However, transmission of massive training data from edge devices to servers is challenging due to limited wireless communication resources, as well as the privacy requirement, which makes it difficult to exploit centralized training for updating the learning model. To solve this problem, federated learning (FL)[4] is proposed, which exchanges learning models rather than raw data between edge devices and edge servers by deploying the training

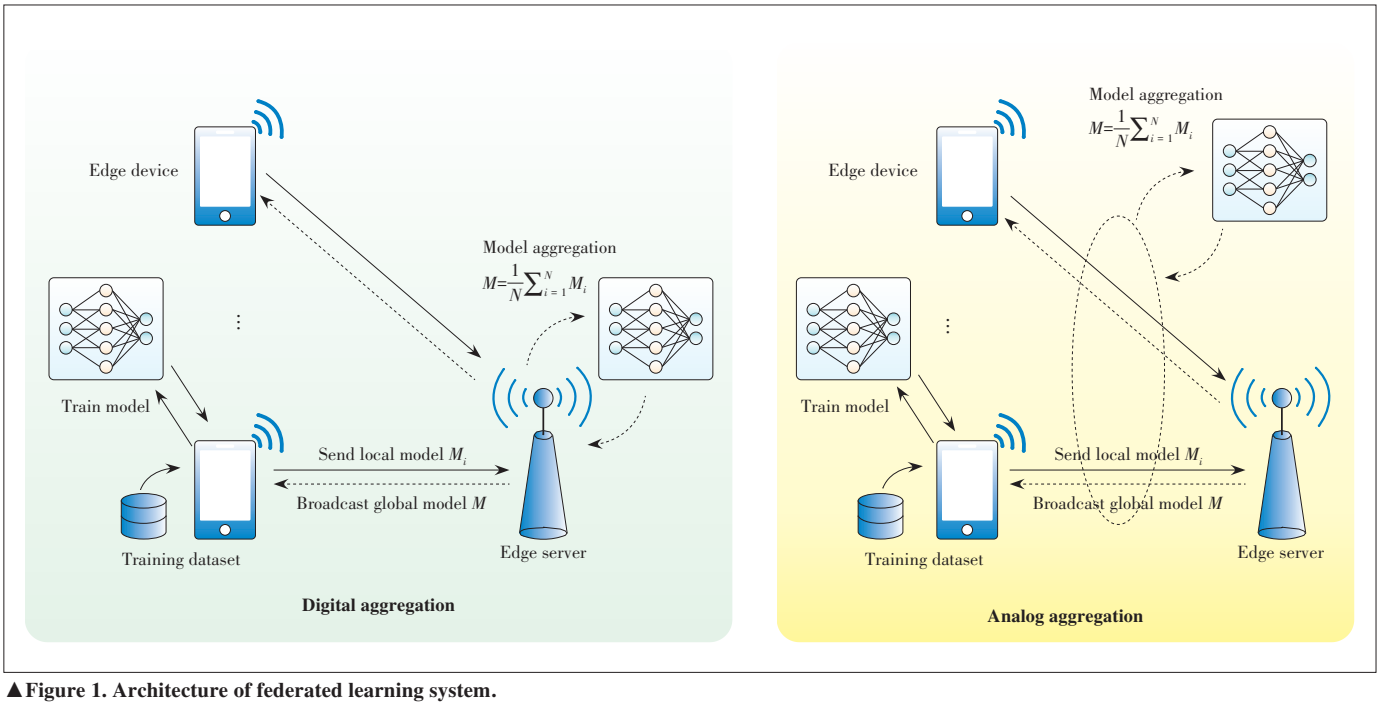
This work is supported in part by the National Key R&D Program of China under Grant No. 2018YFB1800800 and the Nature Science Foundation of China under Grant Nos. 61871254, 91638204 and 61861136003.

algorithms on edge devices. Since mobile devices will consume their limited computation and communication resources when participating in FL, mobile devices may not be willing to contribute. Therefore, some incentives have been introduced, such as the access to the high-quality models trained by FL, as well as some payment after participating in the FL training.

In a typical FL system, there is an edge server and several edge devices, which collaboratively train a learning model. The architecture of FL system is shown in **Fig. 1**. In each iteration (also known as communication round), the edge server aggregates the local models from edge devices in order to update the global model. Then the edge server broadcasts the newest model to edge devices for model training in the next round. After receiving the newest model, each edge device improves this model based on its own data to obtain a new local model. This process goes on until the global model converges. When aggregating the local model, each device can send the gradient of the local model back to the edge server as well as the whole local model. Compared with sending the whole model, sending gradients can reduce the information loss under the constraint of signal-to-noise ratio (SNR) and thus perform better than sending the whole model in analog aggregation (refer to Section 2.2 for details), because the norm of the gradient is smaller than the model generally. Except for analog aggregation, aggregating gradients and models are equivalent from the scheduling point of view, thus we consider that edge devices upload their updated local models rather than model gradients in the following parts of this paper unless otherwise specified. Fig. 1 shows two different model aggregation schemes, analog aggregation and digital aggregation. In the analog aggregation scheme, edge devices send local models to the sever simultaneously and the aggregation is performed in the wireless channel according to the waveform-superposition property. In this way, the system can reduce the transmission latency since the transmission latency will not scale linearly with the number of devices. However, stringent synchronization between devices is needed during the model uploading, and the aggregation is vulnerable under the attack of third-party devices. In the digital aggregation, the model can be encoded for compression, encryption, and other purposes, which prevents the model from being aggregated in the wireless channel and is not suitable for analog aggregation. Although the digital aggregation is more convenient than the analog aggregation, long transmission latency will be introduced when the number of devices is large.

By distributing model training to the edge devices, FL mitigates the problem of privacy leaks caused by sending the raw training data from devices to the server. With the advantage of protecting data privacy, FL has been applied in some data sensitive scenarios, such as health artificial intelligence (AI)[5]. However, some studies show that the learning model can still result in privacy leaks[6]. To solve this problem, differential privacy-based methods[7–8], collaborative training-based methods[9–10] and encryption-based methods[11–12] are proposed, which can protect the privacy of parameters of learning model.

Another advantage of FL is saving the communication cost of transmitting a large amount of training data. However, FL meets some new challenges. The training of the learning model is distributed to edge devices that may have non-indepen-



▲**Figure 1. Architecture of federated learning system.**

dent and identically distributed (non-i.i.d.) training dataset[13], which results in bad performance (such as low accuracy) of the learning model. Also, due to the different computation capabilities of devices, the FL system should consider the synchronization of the model updates from devices, and to address the straggler issues. In practical scenarios, the wireless resources of the FL system are usually limited, and thus the edge server may not be able to receive the local models from all the edge devices. To solve this problem, one direction of research is reducing the cost of transmitting the local model for every edge device, including model compression by quantization[14] and only updating the model for the edge server when the models have significant improvement[15]. Another research direction is the scheduling of devices, where the edge server needs to schedule a subset of edge devices to send the model update. The device scheduling can reduce the communication cost but may result in slower convergence rate of the model training. Given the constrained wireless resources, scheduling policies for FL are proposed to maximize the convergence rate[16] of the learning model or to minimize the energy consumption[17] of the whole system.

There are some existing surveys on FL and edge machine learning[18 - 21]. In Ref. [18], the authors provide a general overview on FL and its challenges in implementation, but do not consider specific issues of deploying FL in wireless networks. The architecture of deep learning and the process of training and inference in the context of edge computing are studied in Ref. [19]. However, the authors of Ref. [19] place more emphasis on optimizing the FL algorithm itself rather than the scheduling policies for FL. The authors of Ref. [20] focus on communication-efficient FL in mobile edge computing platforms, rather than the scheduling policies that maximize the convergence rate of FL under resource constraints. In Ref. [21], the authors discuss potential FL applications in mobile edge computing, the resource allocation problems and data privacy problems in FL. Nevertheless, the authors of Ref. [21] have not provided an in-depth survey on the scheduling policies according to the model aggregation technique of FL in wireless networks, which can greatly affect the design of the scheduling policies.

In summary, none of the existing work has studied the FL in wireless networks from a scheduling perspective. Therefore, we provide a taxonomy on the aggregation methods used in FL, and discuss scheduling policies that can optimize the training performance under resource constraints for both digital-aggregation-based and analog-aggregation-based FL. The rest of this paper is organized as follows. In Section 2, we first introduce FL systems with analog-transmission-based aggregation, and then several scheduling policies designed for the analog-aggregation-based FL are discussed. The scheduling policies designed for the digital-aggregation-based FL are introduced in Section 3. Section 4 gives the conclusion of this paper and the future directions of federated learning.

2 Analog Aggregation

In a conventional wireless system, a base station needs to decode (deliver) the individual information from (to) each user. Accordingly, digital communications and orthogonal multiple access techniques have been developed and widely used. However, a key difference in the FL system is that, while aggregating the local models, the server is not interested in the individual parameters of edge devices, but their average. Note that the waveform-superposition property adds all the signals in a wireless multiple access channel, using analog transmission for global model aggregation, which is a more communication-efficient strategy[22 - 27]. Edge devices synchronize with each other and transmit their local models concurrently. Then the wireless channel carries out the summation over the air, and the server receives the desired values, i.e., the average of the local models, after dividing the received signal by the number of devices involved. Analog aggregation is also called over-the-air computation and it can further support more flexible functions such as weighted summation via power allocation, so that the server can receive the weighted average of local parameters. Some recent papers are summarized in **Table 1**.

2.1 Device Scheduling for Analog Aggregation

A key issue of analog aggregation is how to schedule devices based on their channel states and power constraints. In the t-th round, each device n observes the channel state $h_{n,t}$, and then aligns the transmission power $p_{n,t}$, to ensure that the server can receive its desired value. The power alignment equation is given by

$$p_{n,t}=\begin{cases}\dfrac{a_t}{h_{n,t}}, & \left|h_{n,t}\right|^2>h_{th}\\ 0, & \text{otherwise}\end{cases}, \tag{1}$$

▼**Table 1. Summary of recent papers on analog aggregation**

Technology	Highlights	Related Works
Power alignment	· Fundamental tradeoffs under Rayleigh fading channel	Ref. [23]
	· Online energy-aware dynamic device scheduling policy	Ref. [24]
	· Device scheduling for multi-antenna analog aggregation	Ref. [25]
Sparsification and error accumulation	· Gradient sparsification and error accumulation · Device scheduling policy under average power constraint	Refs. [26 - 27]
Data redundancy	· Introducing data redundancy to deal with non-independent and identically distributed (non-i.i.d.) data	Ref. [24]

where a_t is a power scalar that determines the received SNR at the server side, as well as the energy consumption at the device side. Parameter h_{th} is called the power-truncation threshold, i.e., a device can be scheduled only if its current channel state is better than the threshold. Parameters a_t and h_{th} should be carefully selected in order to optimize the training performance for FL.

In Ref. [23], two fundamental tradeoffs, namely the SNR-truncation tradeoff and reliability-quantity tradeoff, are rigorously characterized. Under the assumption of Rayleigh fading, the relation between the received SNR and power-truncation threshold is studied. The SNR-truncation tradeoff is then revealed: increasing h_{th} can improve the received SNR at the server, at the cost of truncating more devices which cannot satisfy the channel quality requirement. Moreover, the received SNR is limited by the furthest device with largest path-loss. Followed by this observation, a cell-interior scheduling policy is proposed, where only the devices within a distance threshold r_{th} can be scheduled in each communication round. Parameter r_{th} balances the tradeoff between communication reliability and data quantity: larger r_{th} enables the server to schedule more devices and exploit more data for training, while it degrades the received SNR and leads to a noisier version of the average of local models. An alternating scheduling policy is proposed, where the server alternates between the cell-interior scheduling policy and all-included scheduling policy. Finally, theoretical analysis indicates that the communication latency of analog aggregation can be reduced by $O\left(N/\log_2 N\right)$ compared to its digital counterpart, where N is the number of devices.

Removing the Rayleigh fading constraint, an online energy-aware dynamic device scheduling policy is proposed in Ref. [24]. Since the explicit mapping between the loss function of the FL task and the set of devices scheduled in each round remains unknown, an alternative objective function that maximizes the average number of scheduled devices is considered. The long-term average energy constraint (which is equivalent to power constraint) of each device is transformed to a virtual energy deficit queue based on Lyapunov optimization. In each communication round, each device acquires the current channel state $h_{n,t}$ and decides whether to update its local model individually by considering the value of the virtual queue and the required energy consumption. The proposed device scheduling policy works in an online fashion, without requiring any information of the channel states in the future. It also works well if the channel states are non-i.i.d across time.

A multi-antenna analog aggregation FL system is considered in Ref. [25], where the number of scheduled devices is maximized under the mean-square-error (MSE) constraint. Satisfying the MSE requirement can limit the transmission error, and thus it guarantees the accuracy of the aggregated learning model parameters. In order to improve the efficiency of the device scheduling policy, a sparse and low-rank approach is introduced.

2.2 Sparsification and Error Accumulation

The neural networks to be trained for FL tasks usually have huge dimensions, with thousands to millions of parameters. However, the wireless bandwidth is in general limited, and thus the communication latency scales up with the dimension of local models. To further reduce the communication cost for model aggregation, gradient sparsification techniques are introduced in Refs. [26] and [27]. Note that transmitting local gradients rather than local models can improve the power efficiency of analog aggregation, because all the power is used to transmit the information unknown to the server. Therefore, all the devices update their gradients rather than the up-to-date models.

To reduce the dimension of local gradients, a random linear projection is first employed, inspired by compressive sensing. In particular, each local model is multiplied by a random matrix, where each entry follows Gaussian distribution. The random matrix is shared by the devices and the server. Then each device only retains k entries with largest absolute values, which can be regarded as the most important parameters of the gradients, while setting all the other gradients to zero. Here, k is a design parameter which balances the tradeoff between communication reliability and distortion: with smaller k, each entry can be transmitted in a higher power, so that the SNR at the server is higher. However, more information of the local gradient is lost due to the sparsification, degrading the accuracy of the neural network as well as the convergence rate of training.

Instead of discarding all the lost information due to sparsification, a more efficient way is to do error accumulation at the device side. In particular, in each round, the device calculates the differences between the sparse gradients and the original gradients, and adds these differences to the gradients obtained in the next round before employing sparsification. In this way, the error due to sparsification is accumulated by workers, and the training accuracy can be improved according to the experimental results.

Device scheduling policies are also designed for analog aggregation with gradient sparsification and error accumulation. In Ref. [26], additive white Gaussian noise (AWGN) channel is considered, and both equal and unequal power allocation policies are designed. The unequal policy puts more power to the initial rounds, motivated by the fact that the variance of the gradients diminishes across time. Ref. [27] further considers Rayleigh fading channels. Extensive experiments show that compared to the digital aggregation, analog aggregation can improve the convergence rate of training, particularly at low bandwidth and stringent power regimes.

2.3 Non-IID Training Data

The non-i.i.d. data, i.e., the different distributions of data

samples at devices, is also a major bottleneck for FL. It is shown in Ref. [28] that high non-i.i.d. data reduces the accuracy of the neural network by 11% under the Modified National Institute of Standards and Technology (MNIST) dataset, and by over 50% under CIFAR-10 dataset. The non-i.i.d. level of data refers to the difference of local data distribution and global data distribution, which can be characterized by the earth mover's distance, a measurement of the distance between two distributions. To reduce the non-i.i.d. level and thus improve the training accuracy, the server collects some sharable data samples from the devices and disseminates the data to the whole FL system.

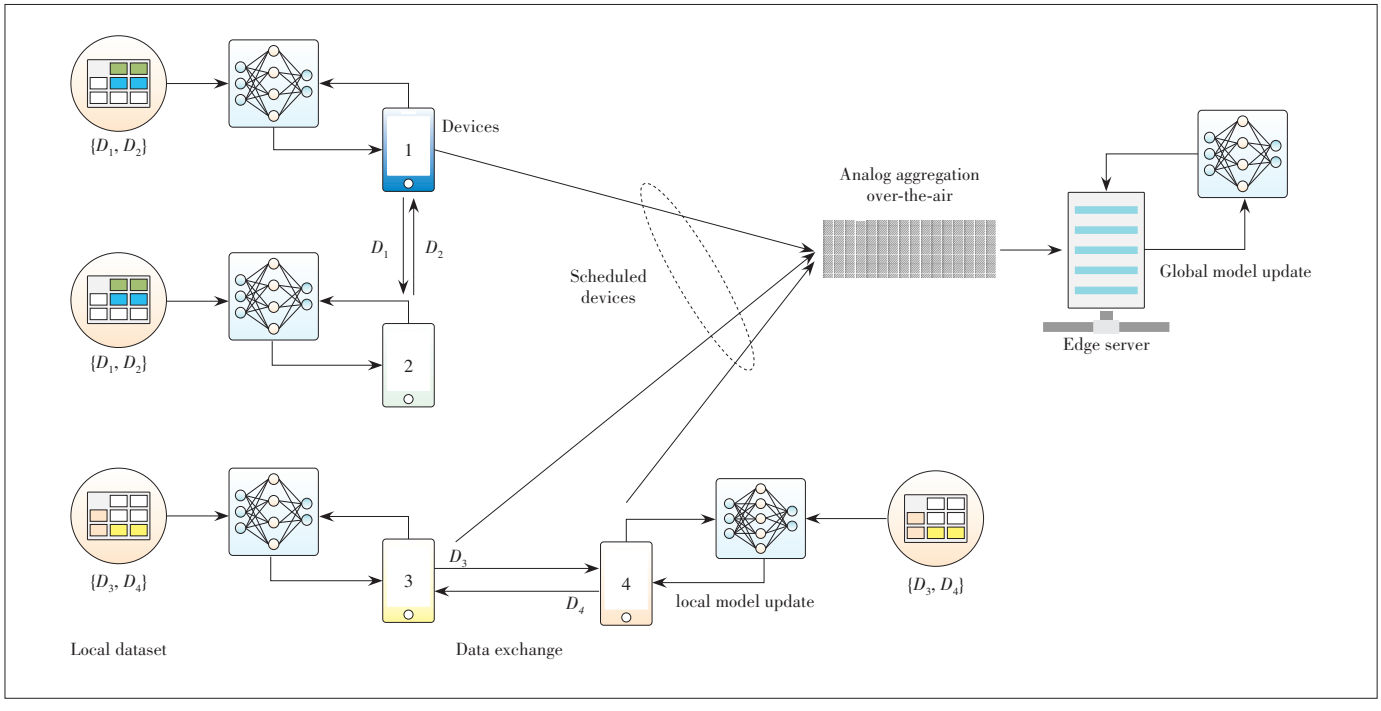
Non-i.i.d. data is still a key issue in analog aggregation FL systems. In Ref. [24], data redundancy is introduced to reduce the non-i.i.d. level of data samples, which can be obtained by exchanging data between a group of devices or collecting data with overlapped coverage in IoT networks. **Fig. 2** illustrates the analog aggregation for FL systems with data redundancy. Workers 1 and 2, 3 and 4 exchange their local datasets with each other, and the redundancy level of the system, i.e., how many devices store each data sample, is two. The experiment results with non-i.i.d. data using MNIST dataset are shown in **Fig. 3**, where $\overline{E}$ is the average energy constraint (in J), "dyn" is the proposed online energy-aware dynamic device scheduling policy, and "myopic" is a benchmark policy where devices can use as much energy as $\overline{E}$ in each round. Parameter r denotes the redundancy level, and $\overline{E} = \infty$ refers to the case where devices have infinite energy, so that all of them can be scheduled in each round. We can see that the proposed dynamic device scheduling policy outperforms the myopic benchmark, and data redundancy can improve the training accuracy significantly. In particular, when $\overline{E} = 5$, increasing redundancy from $r = 1$ to $r = 2$ can achieve an improvement of 10% in training accuracy.

3 Digital Aggregation

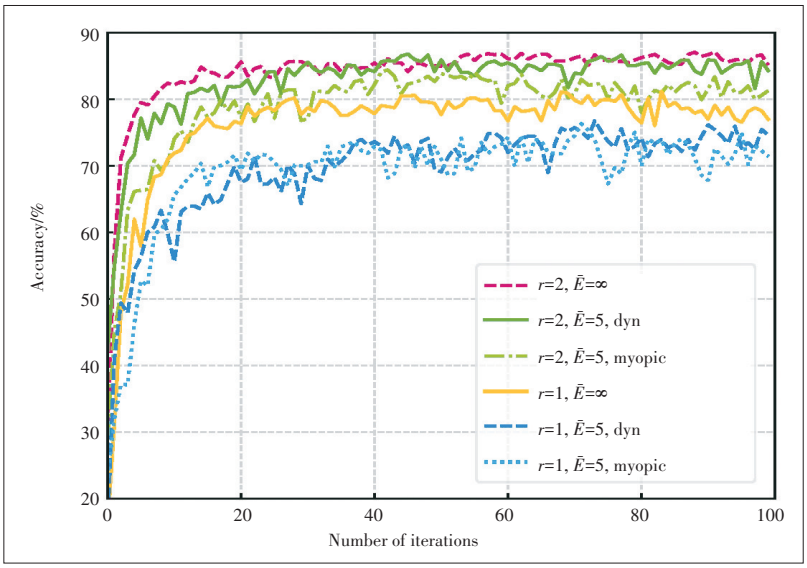
In many other studies, the FL systems are deployed in existing wireless networks (e.g., cellular network or Wi-Fi network), where orthogonal-access schemes such as orthogonal frequency division multiple access (OFDMA) are used for model aggregation. To distinguish them from analog aggregation approaches, we categorize these approaches into digital aggregation. In digital aggregation, the participating devices need to share the scarce wireless bandwidth to upload the updated local models, making the global aggregation very time-consuming. Further, the limited energy and computing resources of participating devices make it more challenging to deploy FL in real wireless networks. Therefore, various scheduling policies have been proposed to address these challenges. These scheduling policies can be divided into the following three categories: aggregation frequency adaptation, local accuracy tuning, and device scheduling. **Table 2** summarizes the highlights of recent papers on digital aggregation.

3.1 Aggregation Frequency Adaptation

In FL, the local update consumes computing resources of devices and the global aggregation consumes the bandwidth resources. Since FL iterates between local updates and global



▲**Figure 2. Analog aggregation for federated learning with data redundancy.**



▲Figure 3. Training accuracy of dynamic device scheduling policy in Ref. [24] under independent and identically distributed (i.i.d.) and non-i.i.d. data.

▼Table 2. Summary of recent papers on digital aggregation

Technology	Highlights	Related Works
Aggregation frequency adaption	· Global aggregation frequency adaption under given resource constraints.	Ref. [29]
	· Extending Ref. [29] into a client-edge-cloud hierarchical FL system	Ref. [30]
Local accuracy tuning	· Tuning local model accuracy to balance the tradeoff between local update and global aggregation · Energy- and convergence-aware resource allocation	Refs. [31 – 32]
Device scheduling	· Energy- and convergence-aware joint scheduling and resource allocation	Ref. [17]
	· Consider unreliable wireless transmissions	Refs. [35 – 36]
	· Maximize the convergence rate with respect to time	Refs. [16] and [37]

aggregations, the frequency of global aggregations (i.e., the reciprocal of the number of local updates between two adjacent global aggregations) should be carefully tuned to balance the consumption of computing and bandwidth resources. In Ref. [29], the authors first analyze the convergence bound of FL with respect to (w.r.t) the number of local updates between two adjacent global aggregations. The bound shows that a higher global aggregation frequency can speed up the FL convergence, while the drawback is consuming more wireless resources for global aggregation. Then a scheduling policy that adapts the frequency of global aggregations in real time to maximize the convergence rate of FL is derived based on the derived convergence bound. The proposed scheduling policy is applicable to non-i.i.d. data distributions and heterogeneous resource constraints of participating devices. Their simulation results show that adaptively adjusting the global aggregation frequency can greatly improve the convergence rate of FL, compared with fixed global aggregation frequency counterparts. Further, the authors of Ref. [30] extend the scheduling policy proposed by Ref. [29] into a client-edge-cloud hierarchical system. In the client-edge-cloud hierarchical FL system, each edge server is allowed to perform partial aggregation that aggregates the updated local models of the edge devices within its communication range. While for the cloud-based global aggregation, the partially aggregated models at edge servers are aggregated through the backbone network by the centralized cloud server. The aggregation frequencies of two levels of model aggregation (i. e., edge-based partial aggregation and cloud-based global aggregation) are optimized to minimize the global loss function value under a constrained number of total local updates.

3.2 Local Accuracy Tuning

The tradeoff between computation and communication is balanced through optimizing the aggregation frequency in aggregation frequency adaptation. Alternatively, some researchers balance this tradeoff via tuning the accuracy level of the local models. In general, increasing local model accuracy requires more computation, while fewer communication rounds are needed for more accurate local models to achieve a fixed global accuracy.

In Ref. [31], the authors refer to an upper bound for the number of communication rounds w.r.t. global accuracy and local accuracy, which is applied to strong convex loss functions for designing the scheduling policy. They adopt time division multiple access (TDMA) for media access control (MAC) layer and dynamic voltage and frequency scaling (DVFS) for devices' CPUs. Thus the frequencies of devices' CPUs, the communication latency of local model uploading and the local accuracy are jointly optimized to minimize the weighted sum of training latency and device energy consumption. As a result, both the computation-communication tradeoff and the device energy consumption-FL training latency tradeoff can be characterized. The overall non-convex optimization problem is decoupled into convex sub-problems, and the closed-form optimal solutions to the sub-problems are illustrated by extensive numerical results. While in Ref. [32], the authors consider a similar FL system but with frequency division multiple access (FDMA). Therefore, the bandwidth allocated to each devices should be jointly optimized with the communication latency, the CPU frequency and the local accuracy. Due to the complicated nature of the problem, the authors of Ref. [32] proposed an iterative algorithm. Their simulation results show that up to 25.6% latency reduction and 37.6% energy reduction can be achieved compared to conventional FL.

3.3 Device Scheduling

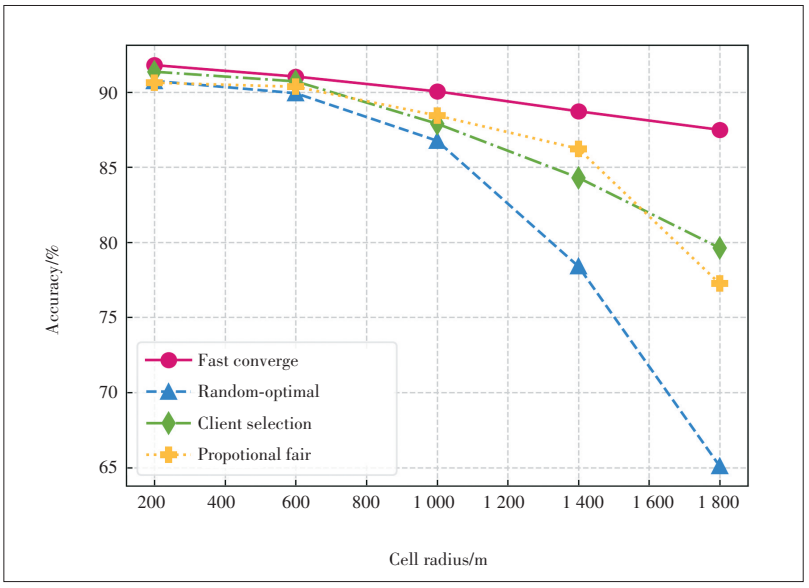
Due to the limited wireless resources and stringent training delay budget, only a portion of devices are allowed to upload local models in each round in real FL systems[33]. Thus the device scheduling policy is critical to FL and will affect the convergence performance in the following two aspects. On one hand, the server needs to wait until all scheduled devices have finished updating and uploading their local models in each round. Therefore, scheduling more devices per round can significantly slow down the model aggregation, because of the reduced bandwidth allocated to each device and the higher probability of having a straggler device (i.e., the device with limited computation capabilities or bad wireless channel). On the other hand, scheduling more devices per round increases the convergence rate (w.r.t. the number of rounds)[34] and can potentially reduce the number of rounds required to attain the same accuracy. To this end, the scheduling policy should carefully balance the latency and learning efficiency per round.

Recently, device scheduling problems in FL have received many research efforts. The authors of Ref. [17] consider a joint scheduling and radio resource allocation problem for FL. In Ref. [17], OFDMA is used for model uploading, where bandwidth allocation can be optimized to reduce the energy consumption. To further characterize the convergence performance, they assume that the convergence rate linearly increases with the number of scheduled devices. Therefore, the optimization objective is set to be the weighted sum of the energy consumption and the number of scheduled devices with a predetermined tradeoff factor, so as to balance the energy consumption and convergence rate. After relaxing the integer constraint for the device scheduling as the real-value constraint, the optimization problem is solved by iteratively solving the bandwidth allocation and scheduling sub-problems.

Furthermore, some recent studies consider the unreliable wireless transmissions. In Ref. [35], the authors propose to deploy FL in cellular networks where inter-cell interference can affect the transmissions of model aggregation. For the transmission quality, only if the received signal-to-interference-plus-noise ratio (SINR) exceeds a threshold, the received local models can be successfully decoded. The convergence rate of FL under such settings, accounting for effects from both scheduling and interference, is then derived. Furthermore, three basic scheduling policies, namely the random scheduling, round-robin and proportional fair, are compared in terms of FL convergence rate. Their results show that the proportional fair policy performs better under a high SINR threshold, while round-robin is suitable for a low SINR threshold. However, the authors of Ref. [36] consider OFDMA for model aggregation and use the packet error rate to capture the unreliability of the wireless transmission. In Ref. [36], a convergence rate bound w.r.t. packet errors is first derived, given the transmitting power of devices, OFDMA resource block allocation and device scheduling policy. Then, the authors formulate an optimization problem to maximize the convergence rate by jointly optimizing the transmitting power allocation, resource block allocation and scheduling policy. The optimization problem is solved in a two-step manner: first obtaining the optimal transmitting power of each device given the device scheduling and resource block allocation; then using the Hungarian algorithm to find the optimal device scheduling and resource block allocation. As shown by simulations, the proposed method can reduce up to 10% and 16% loss function value, compared to: 1) optimal device scheduling with random resource allocation; 2) random device scheduling and random resource allocation, respectively.

However, the convergence rate w.r.t. time, which is critical for real-world FL applications, has not been addressed by aforementioned works. To accelerate the FL training, the authors of Ref. [37] propose to maximize the number of scheduled devices in a given time budget for each round, while the stragglers are discarded to avoid slowing down the model aggregation. The proposed greedy scheduling policy iteratively schedules the device that consumes the least time in model updating and uploading, until reaching the time budget. Although the proposed scheduling policy is simple, their experiments show that it is efficient and applicable to both non-i.i.d. data distributions and heterogeneous devices.

Nevertheless, the time budget is chosen through experiments and can hardly be adjusted under highly-dynamic FL systems. To overcome this drawback, Ref. [16] proposes a joint scheduling and resource allocation policy with fast convergence for FL. Specifically, a latency-optimal bandwidth allocation policy for local model updating and uploading is first derived. Then given the set of scheduled devices and the latency-optimal bandwidth allocation, based on a known upper bound of the number of required rounds to attain a fixed global accuracy, an upper bound of the time required to attain a fixed global accuracy is derived. Finally, an iterative scheduling policy is proposed that iteratively schedules the device that minimizes the approximate time upper bound until the approximate upper bound begins to increase (i.e., scheduling more devices makes the convergence time longer). **Fig. 4** shows the highest achievable accuracy within a total training time budge that equals to 300 seconds under different scheduling policies, including fast converge scheduling policy[16], random scheduling policy with empirically optimal number of scheduled devices (random-opt), client selection policy[37], and proportional fair policy[35]. The experiments are conducted using non-i.i.d. distributed MNIST dataset, and it is assumed that all devices are randomly located in a cell. With different cell radius, the simulation results show that the fast converge scheduling policy always outperforms other scheduling policies in terms of the convergence rate w.r.t. time, and is applicable to non-i.i.d. data.



▲**Figure 4. Highest achievable accuracy under different scheduling policies v.s. the radius of device distributed area.**

4 Conclusions and Future Directions

This paper presents a brief introduction of FL in wireless networks and in particular an overview on the scheduling policies for wireless FL. Firstly, the motivation of deploying FL in wireless networks and the fundamentals of FL systems are introduced. Then, a series of works in the FL systems with analog aggregation are discussed, including device scheduling, model sparsification and data redundancy. Afterwards, we provide an overview on another series of works in FL systems with digital aggregation, including aggregation frequency adaptation, local accuracy tuning and device scheduling. However, apart from the aforementioned works, there are still some challenges and future research directions in deploying FL in wireless networks:

1) Delayed CSI: In the existing works on analog aggregation, power alignment is based on perfect CSIs of devices. While in practice, the server only has delayed CSIs of devices, and how to align the transmission power of devices to minimize the distortion of the aggregated model under delayed CSI remains an open problem. To address this challenge, using the recurrent neural network to predict instantaneous CSI according to the historical CSI estimations may be a future direction.

2) Non-i.i.d. data distribution: Since the data distributions of different devices are usually non-i.i.d. in practical wireless FL applications, it is crucial to design non-i.i.d. data distribution-aware scheduling policies. Although the non-i.i.d. issue in FL systems with digital aggregation has been considered in Refs.[16], [29 – 30] and [37], none of them has proposed any method to alleviate the accuracy degradation caused by non-i.i.d. data. In the future, the data redundancy introduced in Ref. [24] and the communication-efficient data exchange technologies between different devices can be considered in FL systems with digital aggregation to address the non-i.i.d. issue.

3) Convergence guarantee: FL is actually a distributed optimization algorithm that cannot always guarantee to converge. Although most FL algorithms empirically converge and several existing works have provided convergence analysis for FL with convex or strongly convex loss functions. Theoretical analysis and evaluations on the convergence of FL with generally non-convex loss functions are still open problems.

References

[1] CHIANG M, ZHANG T. Fog and IoT: an overview of research opportunities [J]. IEEE internet of things journal, 2016, 3(6): 854 – 864. DOI: 10.1109/JIOT.2016.2584538
[2] ZHU G, LIU D, DU Y, et al. Towards an intelligent edge: wireless communication meets machine learning [EB/OL]. (2018-09-02)[2020-01-31]. https://arxiv.org/abs/1809.00343
[3] PARK J, SAMARAKOON S, BENNIS M, et al. Wireless network intelligence at the edge [J]. Proceedings of the IEEE, 2019, 107(11): 2204 – 2239. DOI: 10.1109/JPROC.2019.2941458
[4] LIM W Y, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey [EB/OL]. (2019-09-26)[2020-01-31]. https://arxiv.org/abs/1909.11875
[5] BRISIMI T S, CHEN R, MELA T, et al. Federated learning of predictive models from federated electronic health records [J]. International journal of medical informatics, 2018, 112: 59 – 67. DOI: 10.1016/j.ijmedinf.2018.01.007
[6] MELIS L, SONG C, CRISTOFARO E DE, et al. Exploiting unintended feature leakage in collaborative learning [EB/OL]. (2018-05-10)[2019-11-01]. https://arxiv.org/abs/1805.04049
[7] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy [C]//Proceedings of 2016 Acm Sigsac Conference on Computer And Communications Security. New York, USA: ACM, 2016: 308 – 318. DOI: 10.1145/2976749.2978318
[8] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective [EB/OL]. (2018-03-01)[2020-01-31]. https://arxiv.org/abs/1712.07557
[9] SHOKRI R, SHMATIKOV V. Privacy-Preserving Deep Learning [C]//Proceedings of 22nd ACM Sigsac Conference on Computer and Communications Security. New York, USA: ACM, 2015: 1310 – 1321. DOI: 10.1145/2810103.2813687
[10] LIU Y, MA Z, MA S, et al. Boosting privately: privacy-preserving federated extreme boosting for mobile crowdsensing [EB/OL]. (2019-07-24) [2020-01-31]. https://arxiv.org/abs/1907.10218
[11] AONO Y, HAYASHI T, WANG L, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. IEEE transactions on information forensics and security, 2017, 13(5): 1333 – 1345. DOI: 10.1109/TIFS.2017.2787987
[12] HAO M, LI H W, XU G W, et al. Towards efficient and privacy-preserving federated deep learning [C]//IEEE International Conference on Communications (ICC). Shanghai, China, 2019: 1 – 6. DOI: 10.1109/icc.2019.8761267
[13] KONEČNÝ J, MCMAHAN B, RAMAGE D. Federated optimization: distributed optimization beyond the datacenter [EB/OL]. (2015-11-11)[2020-01-31]. https://arxiv.org/abs/1511.03575
[14] AJI A F, HEAFIELD K. Sparse communication for distributed gradient descent [C]//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017. DOI: 10.18653/v1/d17-1045

[15] CHEN T Y, GIANNAKIS G B, SUN T, et al. LAG: Lazily aggregated gradient for communication-efficient distributed learning [C]//Adavances in Neural Information Processing Systems 31 (NeurIPS 2018). Montreal, Canada, 2018: 5055 – 5065
[16] SHI W, ZHOU S, NIU Z. Device scheduling with fast convergence for wireless federated learning [EB/OL]. (2019-11-03)[2020-01-31]. https://arxiv.org/abs/1911.00856
[17] ZENG Q, DU Y, LEUNG K K, et al. Energy-efficient radio resource allocation for federated edge learning [EB/OL]. (2019-07-13)[2020-01-31]. https://arxiv.org/abs/1907.06040
[18] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions [EB/OL]. (2019-8-21)[2020-01-31]. https://arxiv.org/abs/1908.07873
[19] WANG X, HAN Y, LEUNG V C M, et al. Convergence of edge computing and deep learning: a comprehensive survey [J]. IEEE communications surveys & tutorials, 2020. DOI: 10.1109/COMST.2020.2970550
[20] SHI Y, YANG K, JIANG T, et al. Communication-efficient edge AI: algorithms and systems [EB/OL]. (2020-02-22). https://arxiv.org/abs/2002.09668
[21] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey [EB/OL]. (2020-02-28). https://arxiv.org/abs/1909.11875
[22] GUNDUZ D, DE KERRET P, SIDIROPOULOS N D, et al. Machine learning in the air [J]. IEEE journal on selected areas in communications, 2019, 37(10): 2184 – 2199. DOI: 10.1109/JSAC.2019.2933969
[23] ZHU G X, WANG Y, HUANG K B. Broadband analog aggregation for low-latency federated edge learning [J]. IEEE transactions on wireless communications, 2020, 19(1): 491 – 506. DOI: 10.1109/TWC.2019.2946245
[24] SUN Y, ZHOU S, GUNDUZ D. Energy-aware analog aggregation for federated learning with redundant data [EB/OL]. (2019-11-01)[2020-01-31]. https://arxiv.org/abs/1911.00188
[25] YANG K, JIANG T, SHI Y, et al. Federated learning via over-the-air computation [EB/OL]. (2019-02-17)[2020-01-31]. https://arxiv.org/abs/1812.11750
[26] AMIRI M M, GUNDUZ D. Machine learning at the wireless edge: distributed stochastic gradient descent over-the-air [C]//IEEE International Symposium on Information Theory (ISIT). Paris, France, 2019: 1432 – 1436. DOI: 10.1109/isit.2019.8849334
[27] AMIRI M M, GUNDUZ D. Federated learning over wireless fading channels [EB/OL]. (2019-07-23)[2020-01-31]. https://arxiv.org/abs/1907.09769
[28] ZHAO Y, LI M, LAI L, et al. Federated learning with non-iid data [EB/OL]. (2018-06-02)[2020-01-31]. https://arxiv.org/abs/1806.00582
[29] WANG S Q, TUOR T, SALONIDIS T, et al. Adaptive federated learning in resource constrained edge computing systems [J]. IEEE journal on selected areas in communications, 2019, 37(6): 1205 – 1221. DOI: 10.1109/jsac.2019.2904348
[30] LIU L, ZHANG J, SONG S H, et al. Edge-assisted hierarchical federated learning with non-iid data [EB/OL]. (2019-10-31)[2020-01-31]. https://arxiv.org/abs/1905.06641
[31] TRAN N H, BAO W, ZOMAYA A, et al. Federated learning over wireless networks: optimization model design and analysis [C]//IEEE Conference on Computer Communications. Paris, France, 2019: 1387 – 1395. DOI: 10.1109/INFOCOM.2019.8737464
[32] YANG Z, CHEN M, SAAD W, et al. Energy efficient federated learning over wireless communication networks [EB/OL]. (2019-11-6)[2020-01-31]. https://arxiv.org/abs/1911.02417
[33] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: system design [EB/OL]. (2019-3-22)[2020-01-31]. https://arxiv.org/abs/
[34] STICH S U. Local SGD converges fast and communicates little [EB/OL]. (2019-05-03)[2020-01-31]. https://arxiv.org/abs/
[35] YANG H H, LIU Z Z, QUEK T Q S, et al. Scheduling policies for federated learning in wireless networks [J]. IEEE transactions on communications, 2020, 68(1): 317 – 333. DOI: 10.1109/tcomm.2019.2944169
[36] CHEN M, YANG Z, SAAD W, et al. A joint learning and communications framework for federated learning over wireless networks [EB/OL]. (2019-9-17)[2020-01-31]. https://arxiv.org/abs/1909.07972
[37] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge [C]//ICC 2019-2019 IEEE International Conference On Communications (ICC). Shanghai, China, 2019: 1-7. DOI: 10.1109/ICC.2019.8761315
[38] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge [C]//IEEE International Conference on Communications (ICC). Shanghai, China, 2019: 1 – 7. DOI: 10.1109/icc.2019.8761315

Biographies

SHI Wenqi received his B.S. degree in electronic engineering from Tsinghua University, China in 2017. He is pursuing his Ph.D. degree in electronic engineering with Tsinghua University. His research interests include edge computing, machine learning and machine learning applications in wireless communications.

SUN Yuxuan received her B.S. degree in telecommunications engineering from Tianjin University, China, in 2015. She is currently working toward the Ph. D. degree in electronic engineering with Tsinghua University. Her research interests include mobile edge computing, vehicular cloud computing and distributed machine learning.

HUANG Xiufeng received his B.S. degree in electronic engineering from Tsinghua University, China, in 2018. He is currently a Ph.D. student in electronic engineering with Tsinghua University. His research interests include machine learning, edge computing and performance optimization for machine learning applications in wireless networks.

ZHOU Sheng (sheng.zhou@tsinghua.edu.cn) received his B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, China, in 2005 and 2011, respectively. He is currently an associate professor of Electronic Engineering Department, Tsinghua University. His research interests include cross-layer design for multiple antenna systems, vehicular networks, mobile edge computing and green wireless communications.

NIU Zhisheng graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992 – 1994, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994, he joined in Tsinghua University, China, where he is now a professor at the Department of Electronic Engineering. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.