As a Ph.D. student in artificial intelligence and machine learning with electrical engineering background, my long-term goal is to build **effective and robust decentralized machine learning (i.e., federated learning** [1, 2, 3] **for the real-world deployment over resource-constrained wireless network**. Federated learning involves computation and communication, where the efficiency can be improved from model training and wireless network perspectives. During my Ph.D. studies, I will approach this goal from three perspectives:

(1) **Improving the convergence rate of federated learning**: One of the new properties that differentiate federated learning from distributed machine learning is the statistical heterogeneity. Particularly, data samples across participating devices may not be independent and identically distributed (non-i.i.d.). Training with non-i.i.d. datasets will lead to biased model updates, stagnating model convergence, and substantially reducing model accuracy. My research analyzed the convergence property given the above-mentioned statistical and/or system heterogeneity from theoretical perspectives, and I further proposed algorithmic and systematic designs that propel faster convergence rate of federated learning.

(2) **Designing lightweight model for on-device training and model transmission**: To alleviate the computation/communication burden caused on devices, I will design the lightweight learning model for on-device local training and model transmission, e.g., the quantization-aware training model. To achieve a massive deployment of FL over resource-constrained devices (e.g., IoT devices in smart manufacturing), I will further design effective resource allocation strategies to vigorously coordinate the candidate devices, thus enhancing the overall system performance.

(3) **Developing the byzantine-resilient federated learning:** Although the raw data are not required to be sent to the cloud server, privacy and security concerns may still arise when the devices and/or the server are curious and malicious (The adversarial devices are referred to as Byzantine devices). Notably, it has been shown that even a single Byzantine fault can significantly alter the trained model with naive mean-value aggregation rule [4]. I will understand and defend different types of adversarial actions, combing with auction-based incentive scheme, to reach a robust federated learning with greater social welfare.

The three tenets of my research—improving the convergence rate of federated learning models, designing lightweight model for resource-constrained network, and byzantine-resilient federated learning—collectively contribute to a future of federated learning that can be deployed over current network infrastructure. I will elaborate more on each of them.

1. Federate Learning with Statistical Heterogeneity

Non-i.i.d is a typical feature when deploying federated learning. For example, Douglas-Fir and Red Maple are distributed in the west part, and east part of the continental United States [5], IoT devices in different locations will collect data samples with different data distribution. To surmount the slow convergence of the most commonly adopted FL algorithm (i.e., FedAvg [1]) under the presence of non-i.i.d. dataset, I analyzed the FL convergence property when heterogeneous data across devices are shown and proposed a weighting strategy[6, 7] and a device scheduling design [8, 9] to improve the convergence rate. Particularly, I observed that

devices with heterogeneous datasets make different contributions to the global model aggregation through examining the gradient information from devices. Therefore, I proposed to assign distinct weights (rather than weigh the local model equally) for participating devices to update the global model. The weighting adaptivity in terms of different devices and different training phases allows the server to accelerate the model convergence. In order to deploy FL over massive devices with non-i.i.d data, partial device participation and device scheduling are needed to be considered. Thus, for the second step, I designed a probabilistic device scheduling framework that can choose devices according to their probability. The probability for each device to be scheduled is determined by the potential contribution from those candidates, and thus is dynamically changed along with the training phases.

The finished works [6, 7, 8, 9] are theoretically proved and empirically verified via extensive experiments.

2. Federate Learning with System Heterogeneity

The complex ML model makes FL deployment on resource-constrained edge devices unrealistic. It is important to consider the lightweight training model to meet the deployable requirements. Preliminary research shows model quantization with a designed number of bit representations achieves competitive training performance (limited number of integer versus float 32) [10, 11]. To reduce the cost of local training and model transmission in the FL context, we can further utilize the concept of model quantization. Particularly, we can adaptively control the quantization level (number of adopted bits) for different parts of the training model (i.e., neural network), given the fact that differentiated quantization sensitivity is shown when quantization is applied on different layers of a neural network (e.g., the convolutional layer is more sensitive compared with fully connected layer when using limited number of integer to represent model parameters), which is testified by our experimental results. Therefore, we will propose effective bit distribution methods (i.e., assign bits to represent different model components) to achieve a larger model representation given the limited communication capability over wireless links.

Due to the large volume of devices in the FL context, resource allocation, e.g., bandwidth allocation, is vital in achieving our third objective. Previous works [12, 13] focus on the resource block (RB) allocation that assigns the limited RBs to the devices with a higher contribution (a larger norm of model update). We can further exploit the layerwise contribution from those devices, given the fact that different layers in the neural networks may behave differently during the training process [14], e.g., the shallow feature in the first several layers may change slightly, contrast to the deep layers. Therefore, we should assign a different amount of communication resource when transmitting the parameters in different layers since those parameters can be quantized with a different number of bits. In addition, given the layerwise contribution from participating nodes, it would be more effective to schedule different devices to transmit different devices of their local updates, i.e., parameters at different layers, to further reduce the communication cost.

My ongoing works focus on this theme, the joint design of model quantization and resource allocation, makes FL deployment over resource-constrained devices realistic.

3. Byzantine-Resilient Federated Learning

Most algorithmic and system designs in federated learning focus on improving the convergence rate or communication reduction without considering the model transmission's reliance. The artificial defacement may occur during model training, i.e., the malicious node, and model transmission, i.e., the noise. The malicious nodes may intentionally affect the local update in many ways, e.g., change the training data, falsify the model parameter, etc., to finally interfere

with the global model [15, 16]. Understanding different types of malicious actions and defending them is also crucial in deploying the FL systems.

The byzantine-resilient federate learning model will also be considered in the context of incentive design, by using auction mechanism[17] to achieve a better social welfare of the federated learning market. Different from the Stackelberg game and contract theory, the auction mechanism allows the data owner to actively report its type. Thus, the FL server can sufficiently understand device status and requests to optimize the target performance metric, such as the social welfare of the market or the server's revenue.

For the next research work, I will focus on the byzantine-resilient federate learning and incentive design to target the robustness feature of federated learning model.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. the Artificial Intelligence and Statistics Conference (AISTATS)*, 2017.

[2] A. Hard, C. M. Kiddon, D. Ramage, F. Beaufays, H. Eichner, K. Rao, R. Mathews, and S. Augenstein, "Federated learning for mobile keyboard prediction," 2018.

[3] H. Ludwig, N. Baracaldo, G. Thomas, Y. Zhou, A. Anwar, S. Rajamoni, Y. Ong, J. Radhakrishnan, A. Verma, M. Sinn, *et al.*, "Ibm federated learning: an enterprise framework white paper v0. 1," *arXiv preprint arXiv:2007.10987*, 2020.

[4] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist species classification and detection dataset," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

[6] H. Wu and P. Wang, "Fast-convergent federated learning with adaptive weighting," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, pp. 1078–1088, 2021.

[7] H. Wu and P. Wang, "Fast-convergent federated learning with adaptive weighting," in *Proc. the IEEE International Conference on Communications (ICC), 2021.*

[8] H. Wu and P. Wang, "Probabilistic node selection for federated learning with heterogeneous data in mobile edge," *accepted to IEEE Wireless Communications and Networking Conference (WCNC)*, 2022.

[9] H. Wu and P. Wang, "Node selection toward faster convergence for federated learning on non-iid data," *accepted to IEEE Transactions on Network Science and Engineering*, 2022.

[10] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Proc. International conference on machine learning*, pp. 1737–1746, PMLR, 2015.

[11] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1709–1720, 2017.

[12] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[13] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2457–2471, 2020.

[14] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4229–4238, 2020.

[15] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *Proc. International Conference on Machine Learning*, pp. 634–643, 2019.

[16] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2168–2181, 2021.

[17] Y. Jiao, P. Wang, D. Niyato, B. Lin, and D. I. Kim, "Toward an automated auction framework for wireless federated learning services market," *IEEE Transactions on Mobile Computing*, vol. 20, no. 10, pp. 3034–3048, 2020.