Primary report about Personalization in Federated Learning

Authors: Mark Zakharov and Rostislav Kulik

1. Problem statement

Personalization in Federated Learning (FL) is a concept that focuses on the optimization formulation of FL. The standard FL aims to find the minimizer of the overall population loss, but this objective has been criticized for many FL applications. The minimizer of the overall population loss might not be the ideal model for a given client, especially if their data distribution differs significantly from the population. Therefore, the problem lies in incorporating personalization into FL to create models that are more suited to individual clients.

2. Importance of personalization in FL

The importance of personalization in FL lies in its potential to significantly improve performance of the model for individual clients. By allowing local models to be mutually different while penalizing their dissimilarity, it's possible to create models that better work with individual clients' data distributions. This can lead to improved accuracy and performance in real-world applications. Furthermore, it can improve communication for problems with heterogeneous data and reduce communication complexity.

3. Examples of occurrence

- A prime example of personalized Federated Learning (FL) models excelling is in the prediction of the next word on a mobile keyboard. Here, a personalized FL approach significantly outperforms a non-personalized one.

- Distributed machine learning provides another instance where data is stored locally across multiple clients. Each client only has access to their own data, which may greatly deviate from the

population average. In such scenarios, personalization can enhance model performance.

4. The main idea of the authors' approach

The authors propose a new optimization formulation for personalized Federated Learning (FL). Instead of finding a single global model that minimizes the overall population loss, they aim to find a mixture of global and local models. This allows each device to learn a personalized model from its own data, while also being influenced by a global model trained on all data.

5. The foundation of the idea

New formulation of FL which seeks an implicit mixture of global and local models

- Zhang, Y. and Yeung, D.-Y. A convex formulation for learning task relationships in multi-task learning.
- Liu, S., Pan, S. J., and Ho, Q. Distributed multi-task relationship learning.
- Wang, W., Wang, J., Kolar, M., and Srebro, N. Distributed stochastic multi-task learning with graph regularization.
- Lan, G., Lee, S., and Zhou, Y. Communication-efficient algorithms for decentralized and stochastic optimization.
- Gorbunov, E., Dvinskikh, D., and Gasnikov, A. Optimal decentralized distributed algorithms for stochastic convex optimization.
- Zhao, P. and Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization.

Theoretical properties of the new formulation

- Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks.

Loopless LGD: non-uniform SGD applied to authors formulation

- Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors.

- Kovalev, D., Horváth, S., and Richtárik, P. Don’t jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop.
- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction.
- Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction.

Convergence theory

- Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates.

Optimal communication complexity and optimal local gradient complexity

- Weiran Wang, Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed stochastic multi-task learning with graph regularization.

Accelerated Gradient Descent

- Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ2).

6. The key features of the proposed methods

- A new optimization formulation for FL that seeks an implicit mixture of global and local models.
- Theoretical properties of the new formulation are studied, developing an algorithm-free theory.
- The proposal of a randomized gradient-based method—Loopless Local Gradient Descent (L2GD)—for solving the new formulation.
- Generalizations allowing for partial participation, local SGD, and variance reduction.

7. Why key features can work well

These methods work well because they address the heterogeneity in data across different devices in FL. By allowing each device to learn a

personalized model, they can better cater to individual data distributions. This can lead to improved accuracy and performance in real-world applications. Furthermore, it can improve communication for problems with heterogeneous data and reduce communication complexity.

8. What are the improvements with the basic versions of what was in the literature before

The improvements over previous methods include providing theoretical convergence guarantees for L2GD, showing that local steps can improve communication for problems with heterogeneous data, and pointing out that personalization yields reduced communication complexity. They also provide generalizations allowing for partial participation, local SGD, and variance reduction. These are significant advancements over traditional FL methods.