جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
ARTIFICIAL INTELLIGENCE INITIATIVE

On Solving a Key Challenge in Federated Learning: Local Steps, Compression and Personalization

Peter Richtárik

KAUST Conference on Artificial Intelligence
April 28-29, 2021

PDF

Filip Hanzely and Peter Richtárik
Federated Learning of a Mixture of Global and Local Models
arXiv:2002.05516, 2020

Part 1
Federated Learning

3 Next Generation AI Technologies

What will the next generation of artificial intelligence look like? Which novel AI approaches will unlock currently unimaginable possibilities in technology and business? This article highlights three emerging areas within AI that are poised to redefine the field—and society—in the years ahead. Study up now.

1. Unsupervised Learning
2. **Federated Learning**
3. Transformers

Forbes

Oct 12, 2020, 09:22pm EDT | 67,594 views

The Next Generation Of Artificial Intelligence

Rob Toews Contributor

AI

I write about the big picture of artificial intelligence.

AI legend Yann LeCun, one of the godfathers of deep learning, sees self-supervised learning as the ... [+]

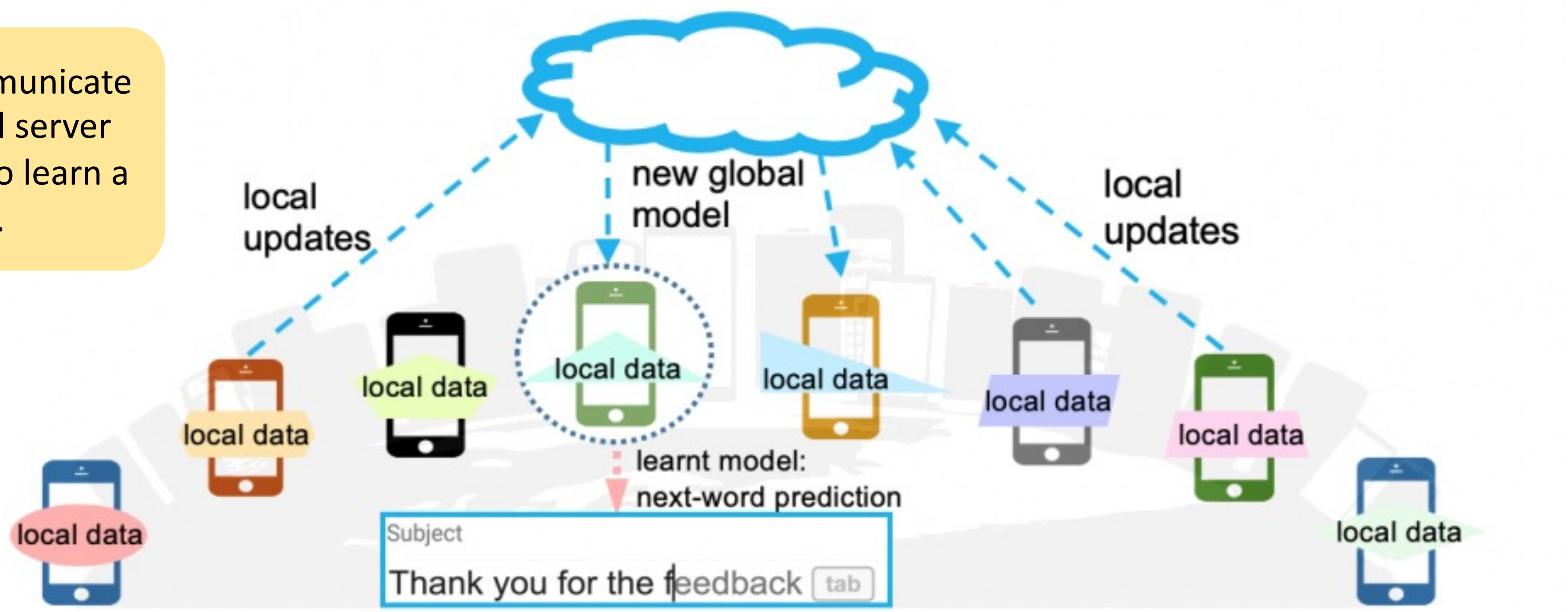
For the second part of this article series, see here.

The field of artificial intelligence moves fast. It has only been 8 years since the modern era of deep learning began at the 2012 ImageNet competition. Progress in the field since then has been breathtaking and relentless.

If anything, this breakneck pace is only accelerating. Five years from now, the field of AI will look very different than it does today. Methods that are currently considered cutting-edge will have become outdated; methods that today are nascent or on the fringes will be mainstream.

Federated Learning: Next Word Prediction

Devices communicate with a central server periodically to learn a global model.

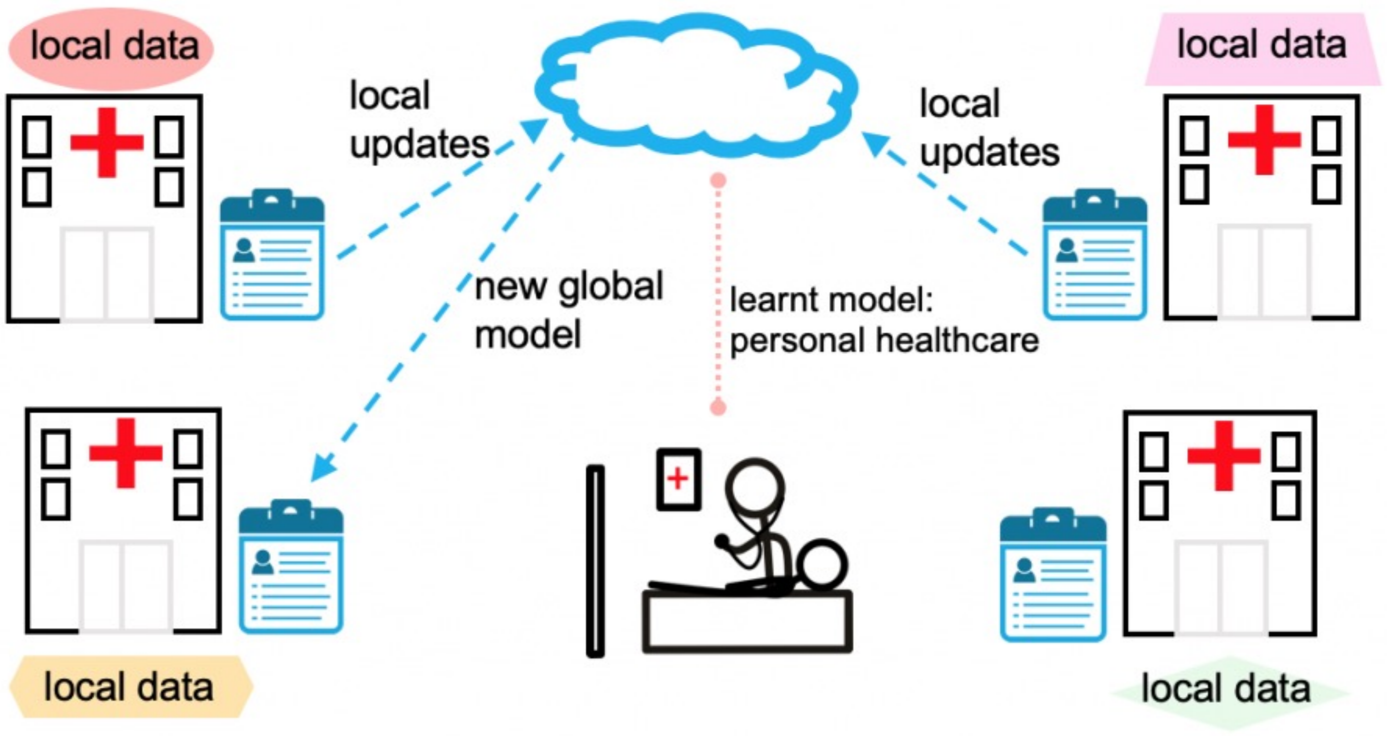


Federated learning helps preserve user privacy and reduce strain on the network by keeping data localized.

https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

Federated Learning: Personalized Healthcare

Devices communicate with a central server periodically to learn a global model.



Federated learning over heterogeneous electronic medical records distributed across multiple hospitals.

https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

First Federated Learning App Launched in 2017

Federated Learning: Collaborative Machine Learning without Centralized Training Data

Google AI Blog

The latest news from Google AI

Thursday, April 6, 2017

Posted by Brendan McMahan and Daniel Ramage, Research Scientists

Standard machine learning approaches require centralizing the training data on one machine or in a datacenter. And Google has built one of the most secure and robust cloud infrastructures for processing this data to make our services better. Now for models trained from user interaction with mobile devices, we're introducing an additional approach: *Federated Learning*.

Federated Learning enables mobile phones to collaboratively learn a shared prediction model while keeping all the training data on device, decoupling the ability to do machine learning from the need to store the data in the cloud. This goes beyond the use of local models that make predictions on mobile devices (like the Mobile Vision API and On-Device Smart Reply) by bringing model *training* to the device as well.

It works like this: your device downloads the current model, improves it by learning from data on your phone, and then summarizes the changes as a small focused update. Only this update to the model is sent to the cloud, using encrypted communication, where it is immediately averaged with other user updates to improve the shared model. All the training data remains on your device, and no individual updates are stored in the cloud.

"We continue to set the pace in machine learning and AI research. ***We introduced a new technique for training deep neural networks on mobile devices called Federated Learning.*** *This technique enables people to run a shared machine learning model, while keeping the underlying data stored locally on mobile phones."*

Sundar Pichai
CEO, Alphabet

4 Foundational Papers Cited in the Blog

4 Foundational Papers of Federated Learning

2016 - 2017

PDF

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson,
Blaise Agüera y Arcas
Communication-Efficient Learning of Deep Networks from Decentralized Data
2/2016

FedAvg algorithm

PDF

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik,
Ananda Theertha Suresh, Dave Bacon
Federated Learning: Strategies for Improving Communication Efficiency
10/2016

Communication compression

PDF

Jakub Konečný, H. Brendan McMahan, Daniel Ramage, Peter Richtárik
Federated Optimization: Distributed Machine Learning for On-Device Intelligence
10/2016

Training via Optimization

PDF

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone,
H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal and Karn Seth
Practical Secure Aggregation for Privacy Preserving Machine Learning
3/2017

Privacy via Secure Aggregation

Jakub Konečný
Research Scientist, Google

I care about making **Federated Learning** happen. Some of that includes research.

I am currently based in Beijing.

Previously, I completed my PhD at the University of Edinburgh under supervision of **Peter Richtárik** and I enjoyed support through **Google PhD Fellowship** in Optimization Algorithms.

White Paper
Financial services
Cyber security
intel
Federated Learning through Revolutionary Technology
A 21st Century Solution for Combating Money
of Terrorism.
TechCrunch
DataFleets keeps private data useful and useful data private with federated learning and $4.5M seed
It hasn't reinvented homomorphic encryption, but has sort of sidestepped it. It uses an approach called federated learning, where instead of ...
Oct 26, 2020
UDACITY
Programs
talog > Secure and Private AI
e and Private AI
by facebook Artificial Intelligence
Learn how to extend PyTorch with the tools necessary to train AI models that preserve user privacy.
ONNECT
Unlock AI, break data silos, and protect privacy with our Federated Learning software
Owkin Connect provides the infrastructure and AI technology that open the possibility for an unprecedented breadth of collaboration in healthcare while protecting patient privacy and data governance. Our distributed architecture and federated learning capabilities allow data scientists to securely
Artificial intelligence / Machine learning
How Apple personalizes Siri without hoovering up your data
The tech giant is using privacy-preservi
improve its voice assistant while keepi
by Karen Hao
9TO5Google
Exclusives
Pixel
Nest
Android
Google TV
Chrom
Google Assistant using federated learni
Android to improve 'Hey Google' ac
Abner Li - Mar. 26th 2021 11:56 am PT
TechCrunch
Sherpa raises $8.5M to expand from conversational AI to B2B privacy-first federated learning services
The turn to building and commercializing federated learning services comes at a time when the conversational AI business found itself stalling.
1 month ago
Hi, how can I help?
TensorFlow Federated: Machine Learning on Decentralized Data
TensorFlow Federated (TFF) is an open-source framework for machine learning and other computations on decentralized data. TFF has been developed to facilitate open research and experimentation with Federated Learning (FL), an approach to machine learning where a shared global model is trained across many participating clients that keep their training data locally. For example, FL has been used to train prediction models for mobile keyboards without uploading sensitive typing data to servers.
TFF enables developers to simulate the included federated learning algorithms on their models and data, as well as to experiment with novel algorithms. Researchers will find starting points and complete examples for many kinds of research. The building blocks provided by TFF can also be used to implement non-learning computations, such as federated analytics. TFF's interfaces are organized in two main layers:
Federated Learning (FL) API
This layer offers a set of high-level interfaces that allow developers to apply the included implementations of federated training and evaluation to their existing TensorFlow models.
import tensorflow as tf
import tensorflow_federated as tff
# Load simulation data.
source, _ = tff.simulation.datasets.emnist
def client_data(n):
return source.create_tf_dataset_for_clie
lambda e: (tf.reshape(e['pixels'], [-
).repeat(10).batch(20)
# Pick a subset of client devices to parti
train_data = [client_data(n) for n in ran
# Wrap a Keras model for use with TFF.
def model_fn():
model = tf.keras.models.Sequential([
tf.keras.layers.Dense(10, tf.nn.
kernel_ini
])
return tff.learning.from_keras_mode

Federated Learning One World Seminar

Weekly on Wednesdays
via Zoom
(41 Talks)

There Won't be a Talk on April 28 Due to a Conflict with KAUST Conference on AI (which you can freely attend in a Zoom webinar form)

Federated Learning One World (FLOW) seminar provides a global online forum for the dissemination of the latest scientific research results in all aspects of federated learning, including distributed optimization, learning algorithms, privacy, cryptography, personalization, communication compression, systems, hardware, and new generation models. The talks will address the theoretical foundations of the field, as well as applications, datasets, benchmarking, software, hardware and systems.

Organizers

- Peter Richtárik, KAUST, Saudi Arabia (Chair)
- Virginia Smith, Carnegie Mellon, USA
- Aurélien Bellet, Inria, France
- Dan Alistarh, IST, Austria

Technical Support

- Samuel Horváth, KAUST, Saudi Arabia

YouTube

Part 2
On Solving a Key Challenge in Federated Learning

Federated Learning of a Mixture of Global and Local Models

Filip Hanzely and Peter Richtárik

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

February 14, 2020

Abstract

We propose a new optimization formulation for training federated learning models. The standard formulation has the form of an empirical risk minimization problem constructed to find a single global model trained from the private data stored across all participating devices. In contrast, our formulation seeks an explicit trade-off between this traditional global model and the local models, which can be learned by each device from its own private data without any communication. Further, we develop several efficient variants of SGD (with and without partial participation and with and without variance reduction) for solving the new formulation and prove communication complexity guarantees. Notably, our methods are similar but not identical to federated averaging / local SGD, thus shedding some light on the essence of the elusive method. In particular, our methods do not perform full averaging steps and instead merely take steps towards averaging. We argue for the benefits of this new paradigm for federated learning.

1 Introduction

With the proliferation of mobile phones, wearable devices, tablets, and smart home devices comes an increase in the volume of data captured and stored on them. This data contains a wealth of potentially useful information to the owners of these devices, and more so if appropriate machine learning models could be trained on the heterogeneous data stored across the network of such devices. The traditional approach involves moving the relevant data to a data center where centralized machine learning techniques can be efficiently applied (Dean et al., 2012; Reddi et al., 2016). However, this approach is not without issues. First, many device users are increasingly sensitive to privacy concerns and prefer their data to never leave their devices. Second, moving data from their place of origin to a centralized location is very inefficient in terms of energy and time.

1.1 Federated learning

Federated learning (FL) (McMahan et al., 2016; Konečný et al., 2016b,a; McMahan et al., 2017) has emerged as an interdisciplinary field focused on addressing these issues by training machine learning

Filip Hanzely and Peter Richtárik
Federated Learning of a Mixture of Global and Local Models
arXiv:2002.05516, February 2020

Training a Federated Learning Model = Solving a Specific Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

d: # model parameters / features

n: # devices

$f_i(x)$: Loss on data $\mathcal{D}_i$ stored on device i

$$f_i(x) = \mathrm{E}_{\xi \sim \mathcal{D}_i} f_\xi(x)$$

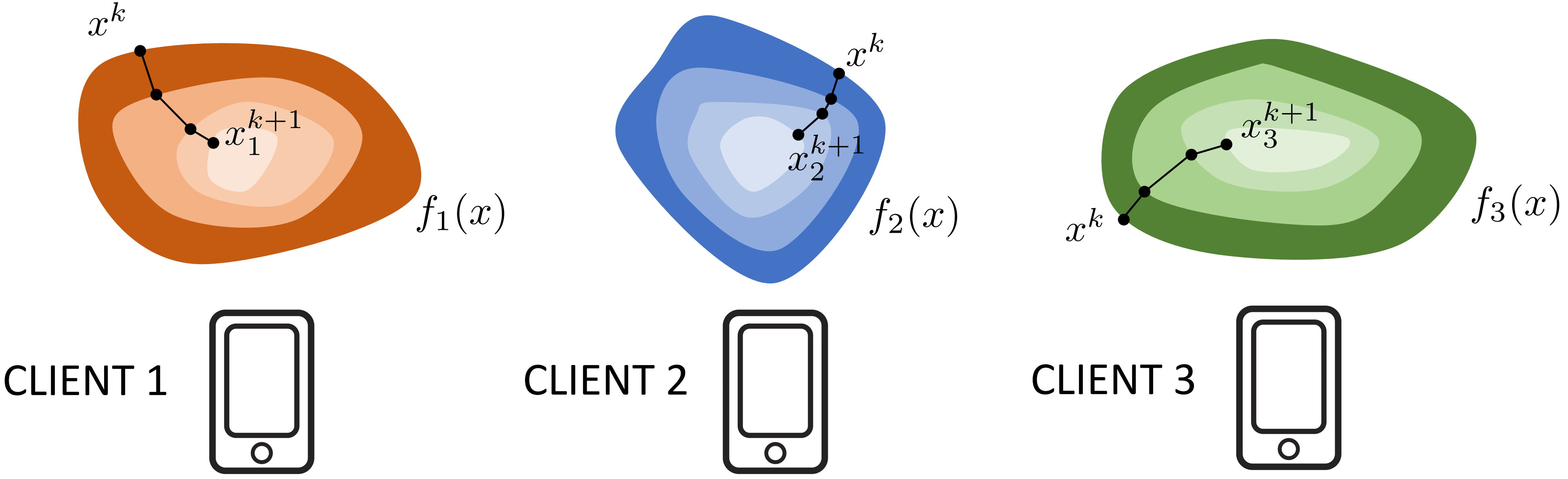
Heterogeneous data regime:

The datasets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_n$ are allowed to be **different**

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas
Communication-Efficient Learning of Deep Networks from Decentralized Data
2/2016

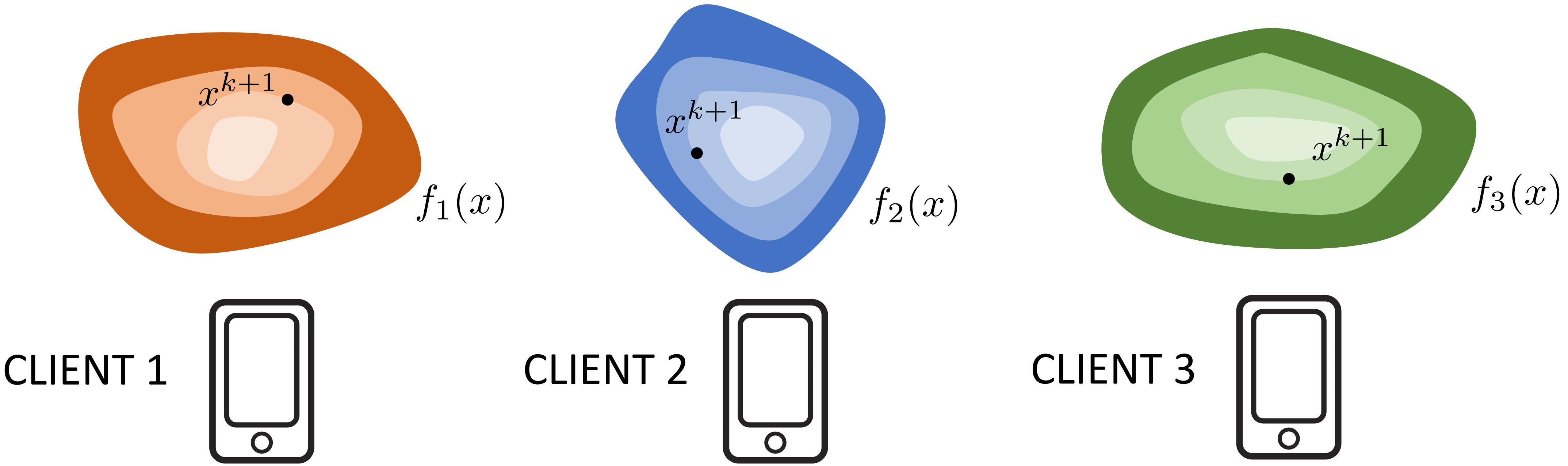
Local Gradient Descent

$$x^{k+1} = \frac{\quad + \quad + \quad}{3}$$



Local Gradient Descent

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas
Communication-Efficient Learning of Deep Networks from Decentralized Data
2/2016



Two Key FL Issues Related to the Heterogeneous Data Regime

model types. Thus, our goal is to use additional computation in order to decrease the number of rounds of communication needed to train a model. There are two primary ways we can add computation: 1) *increased parallelism*, where we use more clients working independently between each communication round; and, 2) *increased computation on each client*, where rather than performing a simple computation like a gradient calculation, each client performs a more complex calculation between each communication round. We investigate both of these approaches, but the speedups we achieve are due primarily to adding more computation on each client, once a minimum level of parallelism over clients is used.

Local Computation

Myth: taking local steps = communication avoidance strategy

Local step methods (such as Local GD) **do not have a theoretical communication complexity advantage** over their non-local counterparts

Ahmed Khaled, Konstantin Mishchenko and P.R.
First analysis of local GD on heterogeneous data
NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality

Ahmed Khaled, Konstantin Mishchenko and P.R.
Tighter theory for local SGD on identical and heterogeneous data
AISTATS 2020

Personalization

There is probably **no single model** that would be good for everyone

Y. Jiang, J. Konečný, K. Rush, and S. Kannan
Improving Federated Learning Personalization via Model Agnostic Meta Learning
arXiv:1909.12488, 2019

We show these issues are connected and propose a solution to both simultaneously!

Our Claims (Very High Level)

1. Local methods for solving ERM can be seen as methods for solving **Personalized ERM (PERM)** instead!

2. When viewed that way, **local methods have (for the forst time!) better communication complexity than nonlocal methods!**

Our New Formulation for FL: Personalized ERM

$x = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{nd}$

$f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$

Regularization parameter $\lambda \geq 0$

$$\min_{x_1, \ldots, x_n \in \mathbb{R}^d} \left\{ F(x) \stackrel{\text{def}}{=} f(x) + \lambda \cdot \psi(x) \right\}$$

$\psi(x) \stackrel{\text{def}}{=} \frac{1}{2n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$

$\bar{x} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i$

Allow different models & penalize dissimilarity

Local GD works well!

Both issues fixed!

Interpolating Two Extremes

$x = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{nd}$

$f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$

Regularization parameter $\lambda \geq 0$

$$\min_{x_1,\dots,x_n \in \mathbb{R}^d} \left\{ F(x) \stackrel{\text{def}}{=} f(x) + \lambda \cdot \psi(x) \right\}$$

$\psi(x) \stackrel{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2$

$\bar{x} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i$

Solution is a function of λ:

$$x(\lambda) = (x_1(\lambda), \dots, x_n(\lambda)) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d = \mathbb{R}^{nd}$$

Local regime $(\lambda = 0)$

$$x_i(0) = \arg\min_{z \in \mathbb{R}^d} f_i(z)$$

No communication is needed!

Global regime $(\lambda = +\infty)$

$$x_i(\infty) = \arg\min_{z \in \mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} f_j(z)$$

$$x_i(\infty) = x_j(\infty) \quad \forall i, j$$

Communication is necessary

Optimality Conditions

$x = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{nd}$

$f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$

Regularization parameter $\lambda \geq 0$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \left\{ F(x) \overset{\text{def}}{=} f(x) + \lambda \cdot \psi(x) \right\}$$

$\psi(x) \overset{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2$

$\bar{x} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i$

Model personalized to device *i*

$$x_i(\lambda) = \bar{x}(\lambda) - \frac{1}{\lambda}\nabla f_i\left(x_i(\lambda)\right)$$

"Meta model" = average of the personalized models

$$\bar{x}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$$

Model Agnostic Meta Learning

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} f_i(z_i) \ : \ z_i = \theta - \alpha \nabla f_i(\theta) \ \forall i \right\}$$

PDF

C Finn, P Abbeel, S Levine

Model-agnostic meta-learning for fast adaptation of deep networks

ICML 2017

Local Methods Developed in our Work

	Random # local steps (L2 = Local & Loopless)	Control variates (i.e., variance reduced method)	On-Device Stochastic Approximatin	Partial Participation of Devices
L2GD	✓			
L2GD+	✓	✓		
L2SGD+	✓	✓	✓	
L2SGD++	✓	✓	✓	✓

L2GD: Loopless Local GD

$x = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{nd}$

$f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$

Regularization parameter $\lambda \geq 0$

$$\min_{x_1,\dots,x_n \in \mathbb{R}^d} \left\{ F(x) \stackrel{\text{def}}{=} f(x) + \lambda \cdot \psi(x) \right\}$$

$\psi(x) \stackrel{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2$

$\bar{x} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i$

Idea: Apply non-uniform SGD to PERM seen as a 2-sum problem!

$$x^{k+1} = x^k - \alpha G\left(x^k\right)$$

Stochastic gradient defined by $\mathbb{E}\left[G\left(x^k\right)\right] = \nabla F\left(x^k\right)$

$$G\left(x^k\right) \stackrel{\text{def}}{=} \begin{cases} \frac{\nabla f\left(x^k\right)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi\left(x^k\right)}{p} & \text{with probability} \quad p \end{cases}$$

Local GD step on all devices

A step towards model averaging

L2GD: Convergence

$$x^{k+1} = x^k - \alpha G(x^k)$$

$$G(x^k) \overset{\text{def}}{=} \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability} \quad p \end{cases}$$

f_i is L-smooth

$$\alpha \leq \frac{1}{2\mathcal{L}} \qquad \mathcal{L} \overset{\text{def}}{=} \frac{1}{n} \max\left\{\frac{L}{1-p}, \frac{\lambda}{p}\right\}$$

$$\mathrm{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \left(1 - \frac{\alpha\mu}{n}\right)^k \left\|x^0 - x(\lambda)\right\|^2 + \frac{2n\alpha\sigma^2}{\mu}$$

On average, $\frac{1-p}{p}$ local steps in between aggregations

On average, $p(1-p)k$ communications per k iterations

f_i is μ-strongly convex

f is $\frac{\mu}{n}$-strongly convex

Optimize over p to minimize number of communications!

$$p^\star = \frac{\lambda}{\lambda + L} \quad \Rightarrow \quad \frac{2\lambda}{\lambda + L} \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \text{communications}$$

PDF R. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin and P.R.
SGD: General Analysis and Improved Rates
ICML 2019

L2GD: # Communications

Local regime $(\lambda = 0)$

$$\frac{2\lambda}{\lambda + L}\frac{L}{\mu}\log\frac{1}{\varepsilon} \to 0 \quad \text{as} \quad \lambda \to 0$$

Global regime $(\lambda = +\infty)$

$$\frac{2\lambda}{\lambda + L}\frac{L}{\mu}\log\frac{1}{\varepsilon} \to \frac{L}{\mu}\log\frac{1}{\varepsilon} \quad \text{as} \quad \lambda \to \infty$$

Rate of GD on ERM

First result for Local GD showing improvement over GD!

Part 3
Adding Compression
for Better Communication Efficiency

Personalized Federated Learning with Communication Compression

El Houcine Bergou [1] Konstantin Burlachenko [1] Aritra Dutta [1] Panagiotis Kalnis [1] Peter Richtárik [1]

Abstract

In contrast to training traditional machine learning (ML) models in data centers, federated learning (FL) trains ML models over local datasets contained on resource-constrained heterogeneous edge devices. Existing FL algorithms aim to learn a single global model for all participating devices, which may not be useful to all devices participating in the training due to the heterogeneity of the data across the devices. Recently, Hanzely and Richtárik (2020) proposed a new formulation for training personalized FL models aimed at balancing the trade-off between the traditional global model and the local models that could be trained by individual devices using their private data only. They derived a new algorithm, called *loopless gradient descent* (L2GD), to solve it and showed that this method leads to improved communication complexity guarantees in regimes when more personalization is required. In this paper, we equip their L2GD algorithm with a *bidirectional* compression mechanism to further reduce the communication bottleneck between the local devices and the server. Unlike other compression-based algorithms used in the FL-setting, our compressed L2GD method operates on a probabilistic communication protocol, where communication does not happen on a fixed schedule. Moreover, our compressed L2GD method maintains a similar convergence rate as vanilla SGD without compression. To empirically validate the efficiency of our algorithm, we perform diverse and numerous numerical experiments on both convex and non-convex problems, and using various compression techniques.

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Correspondence to: any author <name.surname@kaust.edu.sa>.

Preprint, February 2021

1. Introduction

We are living in the era of big data, and mobile devices have become a part of our daily lives. While the training of Machine Learning (ML) models using the diverse data stored on these devices is becoming increasingly popular, the traditional data center based approach to train them faces serious *privacy issues* and has to deal with *high communication and energy cost* associated with the transfer of data from users to the data center (Dean et al., 2012). *Federated learning* (FL) provides an attractive alternative to the traditional approach as it aims to train the models directly on *resource constrained* heterogeneous devices without any need for the data to leave them (Konečný et al., 2016b; Kairouz et al., 2019).

The prevalent paradigm for training FL models is empirical risk minimization, where the aim is to train a *single global model* using the aggregate of all the training data stored across all participating devices. Among the popular algorithms for training FL models for this formulation belong FedAvg (McMahan et al., 2017), Local GD (Khaled et al., 2019; 2020), local SGD Stich (2019); Khaled et al. (2020); Gorbunov et al. (2020a) and Shifted Local SVRG (Gorbunov et al., 2020a) . All these methods require the participating devices to perform a local training procedure (e.g., by taking multiple steps of some optimization algorithm) and subsequently communicate the resulting model to an orchestrating server for aggregation. This process is repeated until a model of suitable qualities is found. For more variants of local methods and further pointers to the literature, we refer the reader to (Gorbunov et al., 2020a).

1.1. Personalized FL

In contrast, Hanzely & Richtárik (2020) recently introduced a new formulation of FL as an alternative to the existing "single-model-suits-all" approach embodied by empirical risk minimization. Their formulation explicitly aims to find a *personalized* model for every device. In particular, Hanzely & Richtárik (2020) considered the formulation[1]

$$\min_{x\in\mathbb{R}^{nd}} [F(x) := f(x) + h(x)] \tag{1}$$

[1]Zhang et al. (2015) considered a similar model in a different context and with different motivations.

El Houcine Bergou
Research Scientist

Konstantin Burlachenko
PhD Student

Panos Kalnis
Professor

Aritra Dutta
Postdoctoral Fellow

PDF

E. H. Bergou, K Burlachenko, A. Dutta, P. Kalnis, and P. Richtárik
Personalized Federated Learning with Communication Compression
Preprint, February 2021

The END