

A Field Guide to Federated Optimization

Jianyu Wang^{*1}, Zachary Charles^{*3}, Zheng Xu^{*3}, Gauri Joshi^{*1}, H. Brendan McMahan^{*3}, Blaise Agüera y Arcas³, Maruan Al-Shedivat¹, Galen Andrew³, Salman Avestimehr¹³, Katharine Daly³, Deepesh Data⁹, Suhas Diggavi⁹, Hubert Eichner³, Advait Gadhikar¹, Zachary Garrett³, Antonious M. Girgis⁹, Filip Hanzely⁸, Andrew Hard³, Chaoyang He¹³, Samuel Horváth⁴, Zhouyuan Huo³, Alex Ingerman³, Martin Jaggi², Tara Javidi¹⁰, Peter Kairouz³, Satyen Kale³, Sai Praneeth Karimireddy², Jakub Konečný³, Sanmi Koyejo¹¹, Tian Li¹, Luyang Liu³, Mehryar Mohri³, Hang Qi³, Sashank J. Reddi³, Peter Richtárik⁴, Karan Singhal³, Virginia Smith¹, Mahdi Soltanolkotabi¹³, Weikang Song³, Ananda Theertha Suresh³, Sebastian U. Stich², Ameet Talwalkar¹, Hongyi Wang¹⁴, Blake Woodworth⁸, Shanshan Wu³, Felix X. Yu³, Honglin Yuan⁶, Manzil Zaheer³, Mi Zhang⁵, Tong Zhang^{3,7}, Chunxiang Zheng³, Chen Zhu¹², and Wennan Zhu³

¹Carnegie Mellon University, ²École Polytechnique Fédérale de Lausanne, ³Google Research, ⁴King Abdullah University of Science and Technology, ⁵Michigan State University, ⁶Stanford University, ⁷The Hong Kong University of Science and Technology, ⁸Toyota Technological Institute at Chicago, ⁹University of California, Los Angeles, ¹⁰University of California, San Diego, ¹¹University of Illinois Urbana-Champaign, ¹²University of Maryland, College Park, ¹³University of Southern California, ¹⁴University of Wisconsin–Madison

Abstract

Federated learning and analytics are a distributed approach for collaboratively learning models (or statistics) from decentralized data, motivated by and designed for privacy protection. The distributed learning process can be formulated as solving federated optimization problems, which emphasize communication efficiency, data heterogeneity, compatibility with privacy and system requirements, and other constraints that are not primary considerations in other problem settings. This paper provides recommendations and guidelines on formulating, designing, evaluating and analyzing federated optimization algorithms through concrete examples and practical implementation, with a focus on conducting effective simulations to infer real-world performance. The goal of this work is not to survey the current literature, but to inspire researchers and practitioners to design federated learning algorithms that can be used in various practical applications.

^{*}Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan conceived, coordinated, and edited this work. Part of the work was done while Jianyu Wang was an intern at Google. Correspondence to Zheng Xu {xuzheng@google.com} and the section editors detailed in acknowledgements and notes.

Contents

1	Introduction	4
1.1	Federated Optimization	5
1.2	Applications	6
1.3	Organization and Background	7
2	Problem Formulation	8
2.1	Federated Optimization Basics	9
2.2	The Generalized Federated Averaging Algorithm	9
2.3	Related Problem Formulations	11
3	Practical Algorithm Design	12
3.1	Guidelines for Developing Practical Algorithms	12
3.1.1	Specify the Application Setting	12
3.1.2	Improve Communication Efficiency	12
3.1.3	Design for Data and Computational Heterogeneity	14
3.1.4	Compatibility with System Architectures and Privacy-Preserving Protocols	15
3.2	Representative Techniques for Improving Performance	16
3.2.1	Incorporate Momentum and Adaptive Methods	16
3.2.2	Reduce the Bias in Local Model Updates	17
3.2.3	Regularize Local Objective Functions	19
3.2.4	Consider Alternative Aggregation Methods	20
4	On Evaluating Federated Optimization Algorithms	21
4.1	Example Evaluations	22
4.2	Suggestions for Evaluations	23
4.2.1	Use Realistic Tuning Strategies	23
4.2.2	Tune Client and Server Learning Rates	24
4.2.3	Analyze Communication-Limited Performance	25
4.2.4	Treat Local Training Steps as a Hyperparameter	26
4.2.5	Understand the Impact of Cohort Size	28
4.2.6	Other Considerations	29
4.3	On the Role of Toy Problems and Artificial Datasets	30
4.3.1	Quadratic and Convex Problems on Synthetic Data	30
4.3.2	Artificial Partitioning of Centralized Datasets	31
4.3.3	Datasets with Examples Shuffled Across Users	32
5	System Constraints and Practices	32
5.1	Communication Costs	32
5.2	Computation and Memory Costs	33
5.3	Analytical Communication and Computation Costs	34
5.4	Basic Model to Estimate Round Time of Cross-Device Training	35
5.5	Research Frameworks for Simulating Distributed Systems	37
5.6	Real-world Deployment Suggestions	38
6	Federated Optimization Theory	39
6.1	Basic Convergence Analysis for Federated Optimization Algorithms	39
6.1.1	Assumptions and Preliminaries	40
6.1.2	Main Results	41
6.1.3	Discussion	42
6.2	Advances in Convergence Guarantees	43

6.2.1	Assumptions on data heterogeneity	43
6.2.2	Extensions and Improvements	44
6.2.3	Gaps Between Theory and Practice	45
7	Privacy, Robustness, Fairness, and Personalization	47
7.1	Privacy	47
7.1.1	Data Minimization	47
7.1.2	Data Anonymization	47
7.2	Robustness	50
7.2.1	Goals of an Adversary	50
7.2.2	Model Poisoning and Data Poisoning Attacks	50
7.2.3	Defenses Against Training-Time Attacks	51
7.2.4	Other Robustness Concerns	52
7.3	Fairness	53
7.4	Tensions Between Privacy, Robustness, and Fairness	53
7.5	Personalization	54
7.5.1	Algorithms that Require Client-side State or Identifier	55
7.5.2	Algorithms that Do Not Require Client-side State or Identifier	55
8	Concluding Remarks	56
A	Datasets	78
A.1	GLD-23k and GLD-160k Datasets	78
A.2	Stack Overflow Dataset	78
A.3	CIFAR-10 Dataset	78
B	Empirical Evaluation - Details	79
B.1	Algorithms	79
B.2	Models	79
C	Additional Experimental Results	80
C.1	More results for Section 4.2	80
C.2	Basic Model to Estimate On-Device Training Times in Section 5.4	84
D	Proofs	86
D.1	Deferred Proof of Lemma 1	86
D.2	Deferred Proof of Lemma 2	87

1 Introduction

Federated learning (FL) was initially proposed by McMahan et al. [181] as an approach to solving learning tasks by a loose federation of mobile devices. However, the underlying idea of training models without collecting raw training data in a single location has proven to be useful in other practical scenarios. This includes, for example, learning from data silos such as businesses or medical institutions which cannot share data due to confidentiality or legal constraints, or applications in edge networks [113, 160, 171, 284]. In light of this, Kairouz et al. [132] proposed a broader definition:

Federated learning is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client’s raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective.

The goal of providing strong privacy protection is implicit in this definition, and a central motivator for FL. Storing data locally rather than replicating it in the cloud decreases the attack surface of the system, and the use of focused and ephemeral updates and early aggregation follow the principle of data minimization. Stronger privacy properties are possible when FL is combined with other technologies such as differential privacy and secure multiparty computation (SMPC) protocols such as secure aggregation. The need to adhere to these privacy principles and ensure compatibility with other privacy technologies puts additional constraints on the design space for federated optimization algorithms.

Federated learning has received increasing attention from both academic researchers and industrial practitioners. The field has exploded from a handful of papers in 2016 including [148, 149, 181], to over 3000 new publications using the term in 2020. Google makes extensive use of federated learning in the Gboard mobile keyboard for applications including next word prediction [104], emoji suggestion [208] and out-of-vocabulary word discovery [53]. Federated learning is also being explored for improving the “Hey Google” detection models in Assistant [95], suggesting replies in Android Messages [91], and improving user experience on Pixel phones [94].

Federated learning is being adopted by industrial practitioners beyond Google. Apple uses federated learning in iOS 13 for applications like the QuickType keyboard and the vocal classifier for “Hey Siri” [12]. Applications of federated learning in the finance space are undertaken by WeBank for money laundering detection [266], as well as by Intel and Consilient for financial fraud detection [126]. In the medical space, federated learning is used by MELLODDY consortium partners for drug discovery [182], by NVIDIA for predicting COVID-19 patients’ oxygen needs [192], by Owkin for medical images analysis [194] and others.

A newer application, federated analytics (FA) [92], poses a class of problems different from the training of deep networks, but nevertheless potentially addressable by optimization methods. As defined in [92], FA is *“the practice of applying data science methods to the analysis of raw data that is stored locally on users’ devices. Like federated learning, it works by running local computations over each device’s data, and only making the aggregated results — and never any data from a particular device — available to product engineers.”* Recent years have seen a dramatic rise in the training of large models from potentially private user data, driving demand for federated learning, but the collection of statistics and use of dashboards for software telemetry is even more ubiquitous, and is fraught with similar privacy concerns. FA thus represents what is likely to become a very common use case, and is already being used in products including Google Health Studies to power privacy-preserving health research [93]. Federated optimization algorithms can be highly relevant to federated analytics: FA can frequently be viewed as training (estimating) a small “model” that is often a simple statistic, such as a count or marginal distribution, rather than a complete distribution

(as in a generative model) or conditional distribution (as in a classifier). Moreover, FA settings can present or accentuate challenges that are relevant to the design of optimization algorithms; we discuss some of these below.

There have been a number of recent works surveying, reviewing, and benchmarking federated learning, including [41, 113, 132, 160, 171, 284]. However, there remains a lack of consensus about core concepts in federated learning, such as the formalization of the optimization problem(s), the definition of data heterogeneity, system constraints, evaluation metrics, relevant experimental settings, and approaches to parameter tuning. This paper provides recommendation and commentary on these areas, among others.

Our goal is also to help bridge the gap between optimization theory and practical simulations, and between simulations and real-world systems. This paper emphasizes the practical constraints and considerations that can motivate the design of new federated optimization algorithms, rather than a specific set of knobs. We do not present new theoretical results, but rather provide suggestions on how to formulate problems in federated optimization, which constraints may be relevant to a given problem, and how to go about empirically analyzing federated optimization. In short, we hope this paper can serve as a guide for researchers designing and evaluating new federated learning algorithms, as well as a concise handbook for federated optimization practitioners.

The goal for this guide is not to be a literature review, a comprehensive survey, nor an empirical benchmark of state-of-the-art methods. In Sections 3 and 6, in addition to the general suggestions, we briefly review advanced techniques in recent research for both practice and theory, and refer readers to related papers for more detailed discussion. The discussion of such techniques is intended to provide representative examples that apply our general guidelines, not to survey all the work in FL. The algorithms are often chosen due to the popularity of the method, as well as authors' familiarity. We encourage readers to focus on the general ideas of how to formulate and analyze FL problems, rather than the specific algorithms discussed.

Disclaimer The practical experience shared in this draft is biased towards the application on mobile devices and some of the data are based on Google's federated learning system [39].

1.1 Federated Optimization

Optimization methods for federated learning must contend with many key issues that centralized methods usually do not have to address; we focus on the common features here, and provide an incomplete list for further discussion in the following sections. Federated learning is interdisciplinary research and some other considerations such as fairness and personalization are also of interest and discussed in Section 2.3 and Section 7.

- *Communication efficiency.* Federated learning is one kind of distributed computing with decentralized data. Minimizing communication between the server and clients is desired both for system efficiency and to support the privacy goals of federated learning. Communication can be the main bottleneck when clients are mobile devices that have limited bandwidth and availability for connection. These needs have led to significant research, including to the development of a plethora of compression mechanisms for reducing communication costs, and improved optimization methods that can reduce the total number of communication rounds. Taking multiple local updates on private data before a global sync (as in federated averaging, see Section 2.1) often achieves surprisingly good results in practice, though a full theoretical characterization of why such algorithms perform well in practice is still lacking.
- *(Data) heterogeneity.* Data is stored across the clients (which typically generate it in the first place), and will never be directly communicated to the server or other clients due to

privacy reasons. The totality of data is typically highly *imbalanced* (clients have different quantities of training data) and *statistically heterogeneous* (the training samples on clients may come from different distributions). In contrast, in conventional distributed optimization settings, a single curator partitions and distributes data among the workers to achieve various goals, such as load balancing, reduction of effective statistical heterogeneity, and so on. The non-IID (identically and independently distributed) data partitioning in federated settings can be challenging when the goal is to train a single global model for all clients, particularly under a limited communication budget.

Aspects of data heterogeneity can be particularly important in federated analytics where online learning or estimation are desirable. In such settings, a small but more time-varying model may be required, as opposed to a possibly static and large one. Analytics applications may require additional statistical rigor in addressing concerns raised from data heterogeneity like estimating confidence intervals for predictions or model parameters, time-of-day variation, and bias from various sources.

- *Computational constraints.* Clients can have different local computation speeds due to inherent hardware differences, or due to competing background tasks, which can lead to challenging computational heterogeneity. If clients are mobile devices or use GPU accelerators, then there may also be strict memory limits for local computation.
- *Privacy and security.* Privacy is an explicit goal in federated learning settings, and the definition of FL itself suggests that algorithms should constrain themselves to only access information from data through aggregates for this reason. Further, the ability to combine federated learning with other privacy technologies, e.g., differential privacy [180] and secure aggregation [38], is of great importance. For many applications, federated optimization methods need to be compatible with such privacy protection methods, or otherwise achieve similar privacy guarantees.
- *System complexity.* Practical FL systems are often complex, e.g., hierarchical aggregation architectures between the coordinating server in the data center and client edge devices. Various protocols are designed to build such complicated systems [39, 199] and they can introduce some implicit constraints for federated optimization. For example, the existence of stragglers (devices that are slow to report updates to the server) can affect algorithmic choices on synchronous or asynchronous optimization. The robustness of the system has to depend on close collaboration of different FL components such as clients for data storage and local computation, communication channels and aggregators for collecting and transferring information, and server for orchestration and global computation.

1.2 Applications

There are many practical applications of the general idea of federated learning. We highlight two settings that may significantly affect the design of practical algorithms: *cross-device FL* and *cross-silo FL*. Cross-device FL targets machine learning across large populations of mobile devices, while cross-silo federated learning targets collaborative learning among several organizations. Both cross-device and cross-silo federated optimization have data and computational heterogeneity as well as communication and privacy constraints, but they also have some nuanced differences [132, Table 1]:

- *Number of clients.* The total number of clients in cross-device FL is typically much larger than in cross-silo FL. The number of clients (organizations) in cross-silo FL could be as small as two, more often dozens or hundreds, while the number of clients (devices) in cross-device FL can easily go beyond millions. The amount of private data on each client in cross-device FL can be smaller than in cross-silo FL. This may also affect the heterogeneity of data, particularly if inherited clustering structures can be found in the large number of clients of cross-device FL.

- *Client availability.* Cross-silo clients are typically hosted in data centers, implying high availability likely enabling all the clients to participate in each round. In contrast, the large and intermittently-available population of typical cross-device settings implies the server can only access clients via some sampling process which can be thought of as given, and the server has only limited ability to control it. In cross-device settings, client availability to participate in training may also correlate with its local training data distribution, introducing diurnal variations that optimization algorithms may need to address [75].
- *Connection topology.* Direct (non-server-mediated) peer-to-peer connections are generally less well-supported on mobile device platforms, whereas in cross-silo settings such direct client-to-client connections may be easier to establish. In cross-silo FL, researchers have explored a variety of communication patterns, e.g., decentralized FL [110, 147, 150, 151, 167, 186, 249, 260], vertical FL [59, 82, 107, 214], split learning [99, 112, 193, 252], and hierarchical FL [40, 170, 254].
- *Computation and communication constraints.* The computation and communication constraints are more strict in cross-device settings. Edge devices often have limited computation and communication power, while clients in cross-silo setting can often be data centers that are able to deploy hardware accelerators and high bandwidth communication links.
- *Client-local computation state.* Given the large populations, client-level privacy goals, and necessity of sampling, in cross-device settings we typically assume devices have no identifier and might only participate once in the entire training process. Thus, in this setting algorithms without client-side state are generally necessary. In contrast, since most clients will participate in most rounds of cross-silo training, stateful algorithms are appropriate.

Note that cross-device and cross-silo settings, mainly defined by their system-imposed constraints, are only two of many possible modalities of federated learning. These two have been the subject of much of the research on federated learning to date, in part due to their immediate application to and practical adoption by industry. Other FL scenarios include networks where ultra-reliable low-latency communication is necessary, such as vehicles on the road [221], or wireless edge networks in general [54, 171, 197].

1.3 Organization and Background

This paper is organized as follows. Section 2 formulates the canonical federated learning problem and introduces the basic optimization framework. Section 3 provides guidelines for developing practical algorithms and presents representative techniques for advanced federated learning algorithm design. Section 4 discusses the use of simulations to evaluate federated optimization algorithms, including concrete examples. Section 5 provides suggestions for deploying FL algorithms on real-world systems. Section 6 reviews the theoretical tools and some recent advances for analyzing FL algorithms. Section 7 draws connections between federated optimization and other important aspects of FL, such as privacy, robustness, personalization and fairness.

Optional background This manuscript is intended to be a self-contained reference on federated optimization and related topics. However, given the wide scope of federated learning and the myriad system-level differences between federated settings, several sections of Kairouz et al. [132] may serve as additional background to the reader who wants more detail: [132, Table 1, Section 1] describes various characteristics and regimes of federated learning; [132, Section 2] discusses applications beyond Section 1.2 including the unique characteristics of *cross-silo* federated learning; [132, Section 3.2] provides an overview of federated optimization methods, theoretical convergence rates, and open problems in federated optimization; [132, Section 4.2] discusses privacy techniques that can be used

in conjunction with federated optimization methods; [132, Sections 5 and 6] provides more discussion of robustness and fairness, and [132, Section 7] introduces system challenges.

2 Problem Formulation

In the basic conception of federated learning, we would like to minimize the objective function,

$$F(\mathbf{x}) = \mathbb{E}_{i \sim \mathcal{P}}[F_i(\mathbf{x})], \quad \text{where} \quad F_i(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi)], \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^d$ represents the parameter for the global model, $F_i : \mathbb{R}^d \rightarrow \mathbb{R}$ denotes the local objective function at client i , and \mathcal{P} denotes a distribution on the population of clients \mathcal{I} . The local loss functions $f_i(\mathbf{x}, \xi)$ are often the same across all clients, but the local data distribution \mathcal{D}_i will often vary, capturing data heterogeneity.

Algorithms designed for the cross-device setting cannot directly compute $F(\mathbf{x})$ or $\nabla F(\mathbf{x})$ because they are assumed to only have access to a random sample \mathcal{S} of the clients in each communication round. However, the objective function $F(\mathbf{x})$ can be used as a mathematical object in the analysis of such an algorithm, or even computed numerically in a simulation as part of an empirical evaluation procedure. If we model cross-device FL with a fixed dataset, to match the population risk (1), we would typically use a held-out set of *clients*¹, rather than a held-out set of examples for each client from a fixed set. We consider this a “stylized” scenario because while such a simplified model of client selection might be suitable for analyzing or comparing some optimization algorithms, other approaches where a client selection strategy is part of the algorithm will require a richer model of device availability and participation.

The cross-silo setting can generally be well-modeled as a finite number of clients, i.e., $F^{\text{silo}}(\mathbf{x}) = \sum_{i=1}^M p_i F_i(\mathbf{x})$. A train/test split can generally be made by splitting the per-client datasets into local train and test sets, with generalization measured against the held-out per-client data.

In both cross-device and cross-silo setting, the objective function in (1) can take the form of an empirical risk minimization (ERM) objective function with finite clients and each client has finite local data:

$$F^{\text{ERM}}(\mathbf{x}) = \sum_{i=1}^M p_i F_i^{\text{ERM}}(\mathbf{x}), \quad \text{where} \quad F_i^{\text{ERM}}(\mathbf{x}) = \frac{1}{|D_i|} \sum_{\xi \in D_i} f_i(\mathbf{x}, \xi) \quad \text{and} \quad \sum_{i=1}^M p_i = 1. \quad (2)$$

Note that $M = |\mathcal{I}|$ denotes the total number of clients and p_i is the relative weight of client i . Setting $p_i = |D_i| / \sum_{i=1}^M |D_i|$ makes the objective function $F^{\text{ERM}}(\mathbf{x})$ equivalent to the empirical risk minimization objective function of the union of all the local datasets. The objective in (1) is equal (in expectation) to the ERM objective that one would optimize centrally if we randomly selected some number of clients and constructed a central training dataset from the union of their local datasets.

Compared to the centralized training, we want to highlight several key properties of (1) and (2):

- **Heterogeneous and imbalanced data:** The local datasets D_i ’s can have different distributions and sizes. As a consequence, the local objectives $F_i(\mathbf{x})$ ’s can be different. For example, they may have arbitrarily different local minima.
- **Data privacy constraints:** The local datasets D_i ’s cannot be shared with the server or shuffled across clients.

¹While in the cross-device setting the server cannot access client IDs to partition clients into disjoint train and test sets, devices can locally flip a coin to decide whether to participate in a training or test population.

- **Limited client availability (more common in cross-device FL):** In cross-device FL, the number of clients M in (2) can be extremely large and not even necessarily well-defined (do we count a device that exists, but never meets the eligibility criteria to participate in training?). At any given time, only a subset (typically less than 1%) of clients are available to connect with the server and participate in the training. The client distribution \mathcal{P} , total number of clients M or the total number of data samples $\sum_{i=1}^M |D_i|$ are not known a priori before the training starts.

2.1 Federated Optimization Basics

Problem (1) can potentially be solved by *gradient descent* (GD), which performs iterations of the form $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \nabla F(\mathbf{x}^{(t)})$, $t = 0, 1, 2, \dots$, where η_t is an appropriately chosen learning rate. Under appropriate regularity conditions, we can swap differentiation and expectation, which gives the following formula for the gradient: $\nabla F(\mathbf{x}) = \nabla \mathbb{E}_{i \sim \mathcal{P}}[F_i(\mathbf{x})] = \mathbb{E}_{i \sim \mathcal{P}}[\nabla F_i(\mathbf{x})]$. Note that the gradient of the global loss function F is equal to the expectation (or “average”) of the gradients of the local functions F_i . In many federated learning settings, the clients can’t communicate among themselves directly, but can communicate indirectly via an orchestrating server. When applying GD to the ERM formulation (2), server has to calculate $\nabla F^{\text{ERM}}(\mathbf{x}) = \sum_{i=1}^M p_i \nabla F_i^{\text{ERM}}(\mathbf{x})$ by weighted average of *all* the local gradients from the clients.

While GD can be conceptually applied in the context of FL, it is not used in practice for various constraints and considerations discussed in Section 1. A number of techniques can be used to enhance GD to make it theoretically or practically efficient as a method for solving federated optimization problems. Moreover, many of these techniques are (orthogonal) enhancements that can be combined for a more dramatic effect. Having said that, many of the possible combinations are not well understood and are still subject of active research.

Partial participation is a requirement for cross-device FL and some cross-silo settings. In communication round t , only a (finite) subset $\mathcal{S}^{(t)}$ of clients can connect to the server, and the update rule becomes $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(\mathbf{x}^{(t)})$. In practice, the sequence of active clients $\mathcal{S}^{(t)}$ is typically dictated by complicated circumstances beyond the control of the orchestrating server (for example mobile devices might only participate when idle, connected to particular unmetered networks, and charging; see Section 5). In theory, assumptions on client sampling are necessary to guarantee convergence (see Section 6 for more discussion), and partial participation can be expected to lead to an increase in the number of communication rounds.

Independently of whether partial participation is necessary, clients can use the *stochastic approximation* (SA), replacing the exact gradient of their local loss function with an unbiased stochastic gradient $g_i(\mathbf{x}^{(t)})$ such that $\mathbb{E}_{\xi \sim \mathcal{D}_i}[g_i(\mathbf{x}^{(t)})] = \nabla F_i(\mathbf{x}^{(t)})$. SA is preferred when the size of D_i is large and the calculation of the exact gradient is inefficient.

Local steps is a popular technique to reduce communication costs. Each active client updates their local model for τ_i steps before the server aggregates the model deltas $\Delta_i^{(t)} = \mathbf{x}_i^{(t, \tau_i)} - \mathbf{x}^{(t)}$. Combining partial participation, stochastic approximation and local steps leads to federated averaging, a popular practical algorithm for federated optimization, which is further discussed in Section 2.2. We defer the discussion of additional techniques like compression, momentum and acceleration, adaptive method, and control variates to Section 3 and the theoretical analysis to Section 6.

2.2 The Generalized Federated Averaging Algorithm

A common algorithm to solve (1) is *federated averaging* (FEDAVG), proposed by McMahan et al. [181]. The algorithm divides the training process into rounds. At the beginning of the t -th round ($t \geq 0$),

the server broadcasts the current global model $\mathbf{x}^{(t)}$ to a *cohort* of participants: a random subset of clients $\mathcal{S}^{(t)}$ (often uniformly sampled without replacement in simulation). Then, each sampled client in the round's cohort performs τ_i local SGD updates on its own local dataset and sends the local model changes $\Delta_i^{(t)} = \mathbf{x}_i^{(t, \tau_i)} - \mathbf{x}^{(t)}$ to the server. Finally, the server uses the aggregated $\Delta_i^{(t)}$ to update the global model:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\sum_{i \in \mathcal{S}^{(t)}} p_i \Delta_i^{(t)}}{\sum_{i \in \mathcal{S}^{(t)}} p_i}, \quad (3)$$

where p_i is the relative weight of client i . The above procedure will repeat until the algorithm converges. In the *cross-silo* setting where all clients participate in training on every round (each cohort is the entire population), we have $\mathcal{S}^{(t)} = \{1, 2, \dots, M\}$.

Algorithm 1: Generalized FEDAVG (also known as FEDOPT [211])

Input: Initial model $\mathbf{x}^{(0)}$; CLIENTOPT, SERVEROPT with learning rate η, η_s

```

1 for  $t \in \{0, 1, \dots, T-1\}$  do
2   Sample a subset  $\mathcal{S}^{(t)}$  of clients
3   for client  $i \in \mathcal{S}^{(t)}$  in parallel do
4     Initialize local model  $\mathbf{x}_i^{(t,0)} = \mathbf{x}^{(t)}$ 
5     for  $k = 0, \dots, \tau_i - 1$  do
6       Compute local stochastic gradient  $g_i(\mathbf{x}_i^{(t,k)})$ 
7       Perform local update  $\mathbf{x}_i^{(t,k+1)} = \text{CLIENTOPT}(\mathbf{x}_i^{(t,k)}, g_i(\mathbf{x}_i^{(t,k)}), \eta, t)$ 
8     end
9     Compute local model changes  $\Delta_i^{(t)} = \mathbf{x}_i^{(t, \tau_i)} - \mathbf{x}_i^{(t,0)}$ 
10  end
11  Aggregate local changes  $\Delta^{(t)} = \sum_{i \in \mathcal{S}^{(t)}} p_i \Delta_i^{(t)} / \sum_{i \in \mathcal{S}^{(t)}} p_i$ 
12  Update global model  $\mathbf{x}^{(t+1)} = \text{SERVEROPT}(\mathbf{x}^{(t)}, -\Delta^{(t)}, \eta_s, t)$ 
13 end
```

FEDAVG can be easily generalized to a flexible framework that allows the algorithm designer to change the client update rule [261, 263, 291], the update rule of the global model [121, 211, 262], or the aggregation method applied to updates [112, 173]. In particular, Reddi et al. [211] proposed a generalized version of FEDAVG, the pseudo-code of which is presented in Algorithm 1. The algorithm is parameterized by two gradient-based optimizers: CLIENTOPT and SERVEROPT with client learning rate η and server learning rate η_s , respectively. While CLIENTOPT is used to update the local models, SERVEROPT treats the negative of aggregated local changes $-\Delta^{(t)}$ as a pseudo-gradient and applies it to the global model. The original FEDAVG algorithm implicitly set SERVEROPT and CLIENTOPT to be SGD, with a fixed server learning rate η_s of 1.0.

The FEDAVG algorithm can be viewed as a generalization of Local SGD (also called local-update SGD or periodic averaging SGD), which is studied for reducing communication cost in classic distributed settings [96, 179, 237, 260, 289]. Some distinguishing properties of FEDAVG are that unlike classic distributed settings, only a subset of clients participate in each training round. In terms of analysis, convergence analyses of Local SGD often assume that the local data is homogeneous and each client performs the same number of local updates, which may not hold in the federated setting.

2.3 Related Problem Formulations

Going beyond the canonical formulation of federated learning of (1), several alternative formulations have been proposed in recent research. We review some such alternative formulations below. This list is admittedly incomplete as federated optimization is rapidly being adopted for myriad new applications. There are several other possible generalizations of the federated learning problem formulation such as 1) federated neural architecture search [111] that searches for efficient neural architecture under FL setting, 2) considering time-varying models $\mathbf{x}(t)$ that adapt to changes to the clients participating in training and their local training data, 3) hierarchical model training, and many more that are beyond the scope of this guide.

Personalization and multi-task learning The data heterogeneity in FL does not only present challenges to the design of provably efficient training methods for solving global objective (1), but also inevitably raises questions about the utility of such a global solution to individual users. Indeed, a global model trained across all the data from all devices might be so different from the typical data and usage patterns experienced by an individual user such that the resulting model becomes worthless. In personalization, every client is allowed to have a different model that is adapted to their local data (i.e., a personalized model). One approach to learning a personalized model is to train a global model and use meta-learning to refine it and obtain personalized models [49, 78, 129, 158]. Another line of work uses multi-task learning [70, 77, 101, 164, 232] to regularize local models towards the global average or towards some reference point. Section 7.5 provides additional detail and discussion of personalized models.

Federated composite optimization Standard FL algorithms such as FEDAVG are primarily geared towards *unconstrained* settings, and f_i 's in (1) are often assumed to be smooth for theoretical analysis. However, many potential FL applications involve non-smooth regularization terms to promote sparsity, low-rankness, monotonicity or enforce hard constraints. Such problems can be formulated as the following *Federated Composite Optimization* (FCO) problem [292]: $\min_{\mathbf{x}} F(\mathbf{x}) + \psi(\mathbf{x})$, with $F(\mathbf{x}) = \mathbb{E}_{i \sim \mathcal{P}}[F_i(\mathbf{x})]$ and $\psi(\mathbf{x})$ a convex, possibly non-smooth, non-finite additive regularizer. See [84, 188] for detailed discussions on the analysis and applications of classic composite optimization problems. To approach the non-smooth (FCO) problem, Yuan et al. [292] propose the *Federated Dual Averaging* (FEDDUALAVG) algorithm, which performs stochastic dual averaging during client local optimization and dual state aggregation during server aggregation.

Probabilistic view Taking a Bayesian perspective [176], we can formulate federated learning as inference of the *global posterior distribution* over the model state \mathbf{x} given the data \mathcal{D} : $\mathbb{P}(\mathbf{x} \mid \mathcal{D} \equiv \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots)$. If the posterior exists, it can be exactly decomposed into a multiplicative average of local posterior distributions (or sub-posteriors): $\mathbb{P}(\mathbf{x} \mid \mathcal{D}) \propto \prod_{i=1}^M \mathbb{P}(\mathbf{x} \mid \mathcal{D}_i)$. Using this decomposition, Al-Shedivat et al. [6] show that the modes of the global posterior correspond to the optima of the objective $F(\mathbf{x})$ and propose to solve federated learning by using a combination of local posterior inference on the clients and global posterior inference on the server. The proposed federated posterior averaging algorithm (FEDPA) resembles generalized FEDAVG (Algorithm 1), has the same computation and communication complexity, but also allows to effectively utilize local computation to de-bias client updates in heterogeneous settings (see discussion in Section 3.2.2). Taking this probabilistic view, Lalitha et al. [156] developed a posterior averaging algorithm to handle non-iid data and general communication graphs where a variational updating of Bayesian posterior density enables a general collaborative learning framework. Unfortunately, however, the variational methods are known to fall short in handling larger models. A handful of recent studies focus on stochastic differential equation (SDE) methods for decentralized sampling [100, 196], making the probabilistic approach more amenable to high dimensions.

3 Practical Algorithm Design

In this section, we will discuss how to design practical federated learning algorithms. We first introduce some general guidelines in Section 3.1 to help readers understand the challenges in designing practical algorithms, and then present several representative algorithms as examples in Section 3.2 that closely follow these guidelines and improve over vanilla FEDAVG. When designing new FL algorithms, readers can use Section 3.1 to check on the practicality and Section 3.2 to get inspirations from previous works.

3.1 Guidelines for Developing Practical Algorithms

3.1.1 Specify the Application Setting

Federated learning can apply to many different applications including the cross-silo and cross-device settings discussed in Section 1.2. These settings have distinct properties such that a single algorithm may not be suitable for all applications. When proposing new optimization algorithms, we recommend researchers clearly specify the most suitable application settings for their algorithms and consider the constraints and requirements of that specific setting.

Stateless clients in cross-device settings In cross-device FL, it is often necessary to design algorithms where clients do not maintain persistent states across rounds. This is due to privacy requirements, but more so due to the fact that populations are typically very large, and so often a particular client will only participate in training a single time. Even in the best case, many rounds are likely to pass between the participations of a single client, and so state would likely be stale and adversely affect convergence.

Large client populations and intermittent availability also imply the server cannot compute any global information that would require an aggregation across all clients. Examples of such global information include the global objective value $F(\mathbf{x})$, the total number of clients M , the total number of samples $\sum_{i \in \mathcal{I}} |\mathcal{D}_i|$, etc. This highlights the fact that the cross-device setting is better modeled by (1), which makes it clear that the above quantities are either ill-defined or can only be estimated.

3.1.2 Improve Communication Efficiency

In federated learning (especially in the cross-device setting), clients may experience severe network latency and bandwidth limitations. Practical FL algorithms generally use communication-reduction mechanisms to achieve a useful computation-to-communication ratio. In this subsection, we are going to introduce three common methods to reduce the communication cost: (i) reduce the communication frequency by allowing local updates; (ii) reduce communication volume by compressing messages; (iii) reduce communication traffic at server by limiting the participating clients per round. These three methods are orthogonal, and can be combined with each other.

Remark 1 (Communication Efficiency). *In this subsection, we consider a method communication-efficient if it can reduce the communication cost per model local update (i.e., per local iteration). However, the total communication cost is a product of number of local iterations and the communication cost per iteration. Therefore, when introducing the communication-efficient methods, we also discuss their impacts on the total number of iterations to achieve a target accuracy.*

Utilize multiple local updates When clients perform τ model updates, the communication cost per client model update can be effectively reduced by a factor of τ . Many works have empirically

validated the effectiveness of local updates in reducing the number of communication rounds to learn an accurate model [47, 172, 179, 181, 185, 202, 233, 241, 259, 296]. While demonstrably effective, this can make analyzing the convergence behavior of methods employing local updates (such as FEDAVG) more difficult than methods such as SGD. In the case of FEDAVG (and Local SGD in many of these works), Stich [237], Wang and Joshi [260], Yu et al. [289] show that there is an additional error term monotonically increasing with the local steps, though it is of a higher order compared to other terms and could be negligible when decaying the learning rate. When client data are non-IID, the side effects of the additional error term will be further exacerbated as shown in [138, 142, 166]. Accordingly, it is worth noting that there is a trade-off between communication reduction and convergence. **When using a larger number of local updates, the algorithm will communicate less frequently and save communication; however, it may incur higher error at convergence** [43, 44]. This trade-off is succinctly captured by the lower bounds for FedAvg in [138, 270]. In practice, one can select a best value of local steps to balance this trade-off or adapt it over time [259].

An alternative to taking multiple local steps is to use a large mini-batch size. Although the local-update SGD approach empirically outperforms the large mini-batch approach [172, 181], its superiority is not well-understood from a theoretical perspective [271]. We refer interested readers to Section 6.1 for more discussion.

Employ compression techniques Reducing the bits transmitted between the clients and the server is another effective approach to save communication costs, and has spurred a large number of methods for sparsifying or quantizing the model updates without substantially reducing accuracy [33, 149, 217, 245, 253, 255, 268]. Compression techniques are especially effective for high-dimensional models. We defer to [281] for a detailed survey of such methods. The output of compression methods can be viewed as stochastic estimates of the model update, and can therefore be *unbiased* or *biased*. Although biased compression methods can achieve higher compression ratios, their direct application can lead to convergence failures [10, 136, 239]. In settings with full participation and where clients have persistent state (*e.g.*, cross-silo FL), these errors can be provably mitigated using error-feedback strategies [30, 35, 97, 116, 136, 146, 174, 205, 206, 225, 238, 240, 247, 280]. Both biased and unbiased compression methods, as well as error feedback schemes, can also be combined with optimization methods, such as momentum, for improved convergence [228, 253]. Another approach incorporates gradient compression into the model architecture itself. For instance, factorized model architectures using low-rank approximation can greatly reduce the amount of communication needed to communicate a model update [149, 258].

Another important facet of compression methods to consider is their compatibility with aggregation protocols in distributed systems. For example, many distributed systems use all-reduce style techniques to aggregate gradients or local model updates across compute nodes [198]. In such systems, compression techniques that are not compatible with all-reduce may provide less communication-efficiency, despite their higher compression ratio [5, 253, 258]. In such settings, it is important to make sure that the compression operation *commutes* with addition. That is, the sum of compressed gradients must be equal (at least in expectation) to the compression of a sum of gradients. This is also necessary for many federated aggregation protocols, such as secure aggregation [38].

Sample a subset of clients Reducing the number of participating clients at each round by sampling a smaller subset can potentially reduce the communication cost as the per-round communication delay monotonically increases with the number of receivers/senders. Moreover, this strategy can mitigate the straggler effects when clients have random or heterogeneous computing speeds [50, 73]. For instance, if we assume that each client’s computing delay is an exponential random variable, then the additional time of waiting for the slowest one is about $\mathcal{O}(\log M)$, where M is the number of participating clients. It is important to note that the number of participating clients influence not

only the communication time but also the convergence properties of the algorithm. Using too few clients per round may significantly increase the stochastic noise in the training process. How to set a proper number of participating clients is still open and less-explored in literature.

Evaluate communication-efficient methods in practice In real-world deployment, the communication efficiency of above methods **should be evaluated via carefully designed wall-clock time measurement**. The wall-clock time of one round of communication will be influenced by algorithmic design choices as well as system properties, including encoding/decoding overhead, fixed network latency (*e.g.*, time to establish handshake that is independent of the number of bits transmitted), and the per-bit communication time multiplied by the model size. Practitioners can choose the most suitable methods according to their specific settings (more discussions can be found in Sections 5.3 and 5.4). For example, compression methods are often evaluated in terms of the total number of bits transmitted. While useful as a rough reference, it does not account for the encoding/decoding times and the fixed network latency in the system. If these delays are much higher than the per-bit communication time, then the benefit of using compression can be incremental or negligible [253]. On the other hand, if the per-bit communication time dominates other delays, the communication savings of compression can be significant. Also, many standard communication stacks will automatically compress network payloads using a general purpose compression scheme like arithmetic coding [215] or the Lempel-Ziv method [302]. So in practice, the gradient/model compression methods are typically used together with standard communication compression algorithms. Understanding the effects of this joint usage is highly useful but still remains under-explored.

3.1.3 Design for Data and Computational Heterogeneity

Due to the imbalance and heterogeneity of client data, the local objectives at clients are generally not identical and may not share common minimizers. Thus, when performing local updates from the same global model, clients will drift towards the minima of local objectives and end up with different local models. This phenomenon is often referred to as *client drift*, and reduces the benefits of performing multiple local steps [e.g. 43, 138]. Detailed discussions on the effects of client drift and its relation to data heterogeneity are given in Section 6.

Another important but less well-studied source of heterogeneity in federated learning is computational heterogeneity. Unlike centralized settings where computations are performed on dedicated and homogeneous machines, clients in federated settings may have varying compute speeds due to hardware differences. For example, some clients may be mobile phones, while others could be laptops equipped with GPUs. Similarly, some clients may dedicate their resources to federated training, while others may need to share resources with background tasks.

While many algorithms assume clients take the same number of steps, this may induce stragglers due to computational heterogeneity, and can increase the runtime of an algorithm. Production systems may enforce a timeout limit, causing the updates from stragglers to be discarded. To avoid such issues, Hsu et al. [122] caps the number of steps any client can take locally, but samples clients with probability proportionally to the amount of data they possess. Another solution is to allow variable number of local steps across clients. For example, Li et al. [161] and Wang et al. [257] propose allowing clients to perform as many local steps as possible within a given time window. Unfortunately, in the presence of data heterogeneity, Wang et al. [261] show that this “variable-step” approach exacerbates client drift and causes a non-vanishing *inconsistency* between the converged point of FEDAVG-style algorithms and the minimizer of actual empirical loss function. They also show that this inconsistency caused by “variable-step” can be removed entirely with careful re-weighting of client updates.

In short, we recommend that researchers and practitioners consider allowing data heterogeneity

and variable amounts of local computation by design. If taking this approach, we recommend modifying the client weighting strategy to account for this variability in steps (*e.g.*, giving less weight to clients which take more local steps [261]). Besides, in some practical scenarios, clients may have very limited computation or memory capacity to finish even one local update. To address this, He et al. [112], Horváth et al. [119] explored a strategy allowing clients use different sizes of models based on their own computation or memory budgets.

3.1.4 Compatibility with System Architectures and Privacy-Preserving Protocols

Practical federated learning systems are often complex and incorporate many privacy-preserving protocols (such as secure aggregation [38] and differential privacy mechanisms). This architectural complexity may introduce additional constraints for practical federated learning algorithms. While it is challenging to design an algorithm compatible with all possible system architectures and privacy-preserving protocols, we recommend that researchers (i) clearly specify the targeted systems and privacy guarantees, (ii) discuss the limitations and incompatibilities of their methodology, and (iii) consider modifications of their proposed algorithms that would allow them to function in different systems. We discuss two examples of additional constraints due to system architectures and privacy-preserving protocols below.

Client inactivity and sampling strategies While cross-device settings may have a large number of clients in the training population, the server may only be able to communicate with a small subset of clients at a given time due to various system constraints. In many experimental simulations, the cohort of clients used to perform training at a given round is obtained by sampling uniformly at random from the total population. In order to derive the theoretical guarantees, it is often assumed that all clients are always available to be sampled. This assumption is often not true in practical cross-device federated learning systems, where clients decide whether to be available based on their own status. For example, a mobile client could only be available when it is connected to WiFi and is charging [39]. This constraint on client sampling is referred to as *client inactivity* in Ruan et al. [218]. Eichner et al. [75] study a related scenario where there are two groups of clients that participate in training at different times of day. For example, users in different geographic locations may only be available at different times of day due to time zone differences (if clients only satisfy the mobile connection criteria at night). This non-uniform client availability leads to optimization challenges with significant practical relevance.

Privacy protection and weighting scheme Privacy protection methods are often applied to the aggregation process (client to server communication) of FL algorithms, for example, secure aggregation [38] and differential privacy [180]. In many variants of federated averaging, the aggregation process can be formulated as a weighted summation of client updates (see equation (3)). A natural choice of the weights for aggregation is the number of examples on clients, so that the expectation of the weighted sum equals the batch gradient for the ERM objective (1) when each client only performs one iteration of gradient update (*i.e.*, FEDSGD). Such sample-based weighting scheme is used in the original FEDAVG algorithm [181] and many other federated learning papers [*e.g.* 166, 261]. However, non-uniform weighting can increase the influence of one client’s update on the aggregated updated, which runs counter to privacy goals (and technically makes bounding sensitivity, a key requirement for differential privacy more challenging). In practice, DP-FEDAVG [180] applies a uniform weighting scheme (*i.e.*, $p_i = 1/|\mathcal{S}^{(t)}|$ in (3)) when applying differential privacy in federated learning. It remains an open question whether uniform weighting by itself can potentially benefit the privacy protection, or robustness to outliers, or fairness to clients with small number of examples.

3.2 Representative Techniques for Improving Performance

In this section we present a few representative techniques and optimization algorithms illustrating how various challenges in federated learning can be addressed. We emphasize that this is not meant to be an exhaustive list of federated optimization algorithms. Instead, the techniques and algorithms mentioned below are chosen to highlight design principles and elucidate settings where they can be applied.

3.2.1 Incorporate Momentum and Adaptive Methods

It is well-known that **momentum and adaptive optimization methods** (such as ADAM [145], ADAGRAD [71], YOGI [209]) have become critical components for training deep neural networks, and can empirically improve the optimization and generalization performance of SGD [204, 246, 298]. A natural question is whether one can carry over the benefits of these methods to the federated setting. To answer this question, many recent papers study how to incorporate momentum and adaptivity in FL, and validate their effectiveness in accelerating convergence.

Server and client optimization framework As discussed in Section 2, Reddi et al. [211] found that vanilla FEDAVG implicitly enables a two-level optimization structure. Accordingly, they proposed a general algorithmic framework named FEDOPT (see Algorithm 1 for the pseudo-code). While clients use CLIENTOPT to minimize the local training loss, the server takes the aggregated local model changes as a pseudo-gradient and uses it as input of SERVEROPT to update the global model. By setting CLIENTOPT or SERVEROPT to be SGDM, ADAM, YOGI, etc., the framework FEDOPT naturally allows the use of any momentum and adaptive optimizers.

Server momentum and adaptive methods The idea of treating local changes as a pseudo-gradient can be traced back to [51] and [189], which apply it to distributed speech recognition and meta-learning, respectively. In the context of federated learning, Hsu et al. [121], Wang et al. [262] show the effectiveness of server-side momentum (*i.e.*, SERVEROPT is SGDM and CLIENTOPT is SGD) both theoretically and empirically. In intuition, the server-side momentum produces a similar effect to increasing the number of selected clients every round. Because of the exponentially weighted averaging of pseudo-gradients, model changes from clients selected in previous rounds also contributes to the global model update in the current round.

In the same spirit, Reddi et al. [211] further demonstrate the benefits of using server-side adaptivity in cross-device federated learning. In particular, by setting ADAM and YOGI as SERVEROPT and SGD as CLIENTOPT respectively, FEDADAM and FEDYOGI incorporate adaptive learning rates into federated optimization, and achieve the much faster convergence and higher validation performance than vanilla FEDAVG on multiple benchmark federated learning datasets. Moreover, the results show that the adaptive methods are easier to tune (*i.e.*, more robust to hyperparameter changes).

Another benefit of using server-side momentum or adaptive methods while keeping CLIENTOPT as SGD is that it does not increase the computational complexity of clients or the communication cost per round. It is also compatible with extremely low client sampling ratio (less than 1%). All the above benefits highlight the utility of server-side momentum or adaptive methods in cross-device FL settings.

Client momentum and adaptive methods Given the server-side momentum and adaptive methods, a natural question that arises is how can we apply momentum or adaptivity directly to the clients? Intuitively, using first- and second-order moments at each step might be better than

using them every τ steps as in server-side adaptive methods. However, when clients locally perform momentum or adaptive optimization methods, their optimizer states are separately updated, and hence, may deviate from each other due to the non-IID data distributions. To address this problem, a few works proposed the synchronized states strategy, where both the local models and the client optimizer states are synchronized by the server at each round. Examples include [288] and [291], both of which apply momentum at clients using the synchronized strategy and assuming all-clients participation. Recently, Wang et al. [263] show that even when the client optimizer states are reset to the default values at the beginning of each round, client adaptive methods can be beneficial, and it is possible to combine with the server adaptive methods to further boost the performance. Besides, this resetting states strategy will not incur any additional communications between the server and clients.

Lazy-update frameworks for momentum and adaptivity Beyond the server and client optimization framework, there also exist other approaches to apply adaptive optimization methods to FL. To avoid deviations among client optimizer states (*i.e.*, first- and second-order moments in adaptive optimizers), a common idea is to fix the optimizer states during local training and lazily update them on the server at the end of each round. Compared to the server-side adaptive methods mentioned above, there are two key differences: (i) The lazy-update algorithms allow the clients to locally use momentum or adaptive methods, but their optimizer states are fixed or keep synchronized; (ii) In these lazy-update algorithms, the optimizer states are updated on the server either by synchronizing all clients’ local optimizer states or by using the batch gradient evaluated on the current global model.

For instance, Xie et al. [279] proposed Local ADAALTER, which combines Local SGD with ADAGRAD via lazy updates of the optimizer states. In local ADAALTER, the clients use ADAGRAD instead of SGD to perform local updates. However, the optimizer states at different clients are always the same at each local iteration, as they are updated in the exact same way without any local information. This algorithm is designed for the setting where all clients can participate in every round of training and its effectiveness has been evaluated on classic distributed training tasks. Moreover, in order to synchronize the clients’ optimizer states, it requires doubled communication cost compared to FEDAVG and the server-side adaptive methods.

A similar idea of using lazily updated optimizer states at clients appears in [137]. The authors proposed the MIME framework to adapt centralized stateful optimization methods to cross-device FL settings. The proposed algorithms split the model update and optimizer states update between the server and the clients. In particular, at each round, the server broadcast the global model \mathbf{x} and the optimizer states \mathbf{s} to a random subset of clients. Then, the selected clients perform local updates according to the update rule of momentum or adaptive optimizer, but their optimizer states \mathbf{s} remain unchanged during local training. Finally, the server uses the aggregated local model changes and the synchronized local batch gradient $\frac{1}{|S|} \sum_{i \in S} \nabla F_i(\mathbf{x})$ to update the global model and optimizer states, respectively. The above procedure repeats until convergence. It is worth noting that Karimireddy et al. [137] showed that using momentum in this way, along with an additional *control variates* technique, can help to reduce the client drift caused by taking local steps. Furthermore, compared to server-side adaptive method (with a fixed server learning rate in simulation), MIME (and its more practical variant, MIMELITE) can achieve better or comparable performance on several federated training tasks, albeit at the expense of additional communication and computation costs.

3.2.2 Reduce the Bias in Local Model Updates

As mentioned, FEDAVG and other federated optimization algorithms generally reduce the communication cost by allowing clients to take multiple SGD steps per round, before synchronizing with the

server. However, in settings with heterogeneous client data, taking more local steps in fact hinders convergence because the resulting client updates ($\Delta_i^{(t)}$'s in Algorithm 1) become biased towards the local minimizers. Here, we introduce two methods that can help to reduce the bias of each client's local model updates.

Control variates *Control variates* is a technique developed in standard convex optimization literature to reduce the variance of stochastic gradients in finite sum minimization problems and thereby speed up convergence. It is possible to adapt the technique to the federated learning setting to *reduce the variance across clients*. A representative algorithm employing this technique in FL is SCAFFOLD [138], as we describe below.

The SCAFFOLD algorithm is applicable to the *cross-silo* setting and makes use of a persistent state stored with each client. This state is a control variate \mathbf{c}_i for client i which is meant to estimate the gradient of the loss with respect to the client's local data $\mathbf{c}_i \approx \nabla F_i(\mathbf{x})$. The server maintains the average of all client states as its own control variate, \mathbf{c} , which is communicated to all the selected clients in each round. Clients perform multiple steps of SGD in each round, just like FEDAVG, but adding a correction term $\mathbf{c} - \mathbf{c}_i$ to each stochastic gradient. The effect of the correction is to *de-bias* the update step on each client, ensuring they are much closer to the global update $\nabla F_i(\mathbf{x}) + \mathbf{c} - \mathbf{c}_i \approx \nabla F(\mathbf{x})$. This enables SCAFFOLD to provably converge faster than vanilla FEDAVG without any assumptions on the data heterogeneity.

When implementing SCAFFOLD, there are many possible choices of the control variates. For example, one can choose to use the averaged local gradients in the last round as \mathbf{c}_i . Specifically, after local updates, the local control variate is updated as follows:

$$\mathbf{c}_i^{(t+1)} = \begin{cases} \mathbf{c}_i^{(t)} - \mathbf{c}^{(t)} + \frac{1}{\eta\tau}(\mathbf{x}^{(t)} - \mathbf{x}_i^{(t,\tau)}), & \text{if } i \in \mathcal{S}^{(t)} \\ \mathbf{c}_i^{(t)}, & \text{otherwise} \end{cases} \quad (4)$$

where the superscript (t) denotes the index of communication round. This strategy is used in [138]. It also appears in [169] when assuming all-clients participation. It is worth noting that (4) requires clients to have persistent states across rounds. In order to overcome this limitation, the MIME algorithm [137] explores another option, that is, using the stochastic gradient evaluated on the global model $\nabla F_i(\mathbf{x}; \xi_i)$ as the local variate \mathbf{c}_i and the synchronized full-batch gradient $\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla F_i(\mathbf{x})$ as the global control variate \mathbf{c} . By doing this, MIME can also reduce the variance in local mini-batch gradients and is applicable to the cross-device setting.

Local posterior sampling Another way to de-bias client updates is computing client deltas as follows [6]:

$$\hat{\Delta}_i^{(t)} = \hat{\Sigma}_i^{-1}(\hat{\mu}_i - \mathbf{x}^{(t)}), \quad (5)$$

where $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are estimates of the mean and covariance of the local posterior distribution on the i -th client, $p(\mathbf{x} | \mathcal{D}_i)$. As the number of posterior samples used to estimate $\hat{\mu}_i$ and $\hat{\Sigma}_i$ increases, the bias in $\hat{\Delta}_i^{(t)}$ vanishes. In other words, when clients can use some extra compute per round, instead of running SGD for many steps or epochs (which would result in biased deltas in heterogeneous settings) we can instead run local Markov chain Monte Carlo (MCMC) to produce approximate local posterior samples and use them to reduce the bias. Under the uniform prior assumption, the global posterior can be estimated by the local posteriors.

Al-Shedivat et al. [6] designed FEDPA algorithm that allows to compute $\hat{\Delta}_i^{(t)}$ efficiently, by adding only a small constant factor computational overhead and no communication overhead compared to FEDAVG, and used it as a drop-in replacement in Algorithm 1. FEDPA works with stateless

clients, and is applicable for cross-device FL. Methods similar to FEDPA have been developed in the distributed inference literature, before FL was introduced, such as consensus Monte Carlo [224] and expectation propagation (EP) [108, 251]. In fact, EP can be seen as a generalization of FEDPA, although it requires stateful clients, and hence can be used only in cross-silo settings. The generalization to arbitrary graphs for collaborative learning was discussed in Lalitha et al. [156] and Nedić et al. [187]; however, the challenges of communicating posteriors efficiently and accurately remain largely unaddressed.

3.2.3 Regularize Local Objective Functions

In vanilla FEDAVG and FEDOPT, clients perform multiple local updates at each round. As discussed in Section 3.1.2, while more local steps will reduce the average communication delay per iteration, it may incur additional errors at the end of training due to the heterogeneous local objectives. If we let clients perform too many local steps, then the local models will directly converge to the minima of the local objectives, which can be far away from the global one. In order to avoid local models drift towards their local minima, a natural idea is to penalize local models that are far away from the global model by regularizing the local objectives. Formally, one can augment the local objective function at the t -th round as follows:

$$\tilde{F}_i^{(t)}(\mathbf{x}) = F_i(\mathbf{x}) + \psi_i(\mathbf{x}, \mathbf{x}^{(t)}) \quad (6)$$

where $\mathbf{x}^{(t)}$ is the global model and also the starting point of local training at round t , and function ψ_i denotes a kind of distance measure between the local model \mathbf{x} and the current global model $\mathbf{x}^{(t)}$. By carefully choosing the regularizer ψ_i , we can ensure the local models are close to the global one, or the surrogate local objectives (6) have the same stationary points as the original global objective (1). Some examples of ψ_i in literature are listed below.

Example: FEDPROX [161]. This algorithm uses the Euclidean distance between local models and the global model as the regularization function. That is, $\psi_i(\mathbf{x}, \mathbf{x}^{(t)}) = \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^{(t)}\|^2$ where $\mu \geq 0$ is a tunable hyperparameter. This easy form of regularization is particularly applicable to the cross-device setting of FL and does not need any additional computation or communication compared to vanilla FEDAVG. Nonetheless, in theory, it cannot offer better rate of convergence than vanilla FEDAVG, see the discussions and analyses in [43, 138, 261].

Example: FEDDANE [159]. Inspired by the DANE algorithm in classic distributed optimization, Li et al. [159] further add a linear term to the regularization function as follows:

$$\psi_i(\mathbf{x}, \mathbf{x}^{(t)}) = \left\langle \frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(\mathbf{x}^{(t)}) - \nabla F_i(\mathbf{x}^{(t)}), \mathbf{x} - \mathbf{x}^{(t)} \right\rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^{(t)}\|^2. \quad (7)$$

In order to compute the linear term (inner product) in (7), the selected clients need to compute the full-batch local gradients and synchronize them, resulting in doubled communication cost per round. Also, as noted by the authors, despite encouraging theoretical results, FEDDANE demonstrates underwhelming empirical performance compared to FEDPROX and FEDAVG. In certain cases (*e.g.*, client optimizer is SGD and $\mu = 0$), FEDDANE has a similar client-side update rule to MIME [137]. Specifically, FEDDANE adds a correction term $\frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(\mathbf{x}^{(t)}) - \nabla F_i(\mathbf{x}^{(t)})$ to each local stochastic gradient, while MIME uses $\frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(\mathbf{x}^{(t)}) - \nabla F_i(\mathbf{x}^{(t)}; \xi_i^{(t,k)})$ where $\xi_i^{(t,k)}$ denotes a random mini-batch sampled at local iteration k and round t .

Example: FEDPD [299] and FEDDYN [2]. These two algorithms similarly modify the local objectives, and define the regularizer of local objective at client i as follows:

$$\psi_i(\mathbf{x}, \mathbf{x}^{(t)}) = \langle \boldsymbol{\lambda}_i^{(t)}, \mathbf{x} - \mathbf{x}^{(t)} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^{(t)}\|^2 \quad (8)$$

where $\boldsymbol{\lambda}_i^{(t)}$ is an auxiliary variable and updates at the end of each round: $\boldsymbol{\lambda}_i^{(t+1)} = \boldsymbol{\lambda}_i^{(t)} + \mu(\mathbf{x}_i^{(t, \tau_i)} - \mathbf{x}^{(t)})$. By doing this, Acar et al. [2] proves that the surrogate local objectives, in the limit, have the same stationary points as the global objective. As a consequence, when the surrogate local functions are minimized exactly at each round, FEDDYN can achieve faster convergence rate than FEDAVG for convex smooth functions. Although the regularization (8) does not introduce extra communications, it requires clients to maintain persistent states or memory across rounds. Hence, it is more suitable for cross-silo FL settings.

3.2.4 Consider Alternative Aggregation Methods

As noted earlier, the global objective $F(\mathbf{x})$ defines a weighting over the individual client objectives $F_i(\mathbf{x})$. Typically, clients' local updates may be weighted by number of examples or weighted uniformly at each round. Though the weighting schemes may not influence the convergence rate of FL algorithms, they do control where the global model finally converges to. By tuning the weighting scheme, **it is possible that the final global model is biased towards the minima of certain clients.** We illustrate these effects in detail below.

Objective inconsistency problem When all clients participate in training at each round and have homogeneous computing speeds, the server just needs to aggregate all local changes in the same way as the global objective. However, when there is client sampling mechanism as in cross-device FL settings or clients take different local steps per round, the weighting scheme closely relates to many implementation choices. If the weighting scheme is not properly chosen, then the global model will instead converge to the stationary points of a surrogate objective, which is inconsistent with the original one.

For example, Li et al. [166] show that, in order to guarantee convergence, the weighting scheme should be selected according to the client sampling scheme. Specifically, if clients are sampled with replacement based on their local sample sizes, then the local model changes should be uniformly averaged; if clients are uniformly sampled without replacement, then the weights of each local model should be re-weighted according to their sample sizes. *The key is that we need to ensure the expectation of aggregated local changes weights each local client in the same way as the global objective.* Otherwise, there is a non-vanishing term in the error upper bound, which does not approach to zero even if learning rate is small enough or gradually decayed [60]. Researchers often assume weighted sampling with replacement and uniform averaging for the simplicity of theoretical analysis. However, in real-world cross-device FL settings, the server has both limited knowledge of the population and limited control over clients sampling, and so weight in aggregation ends up effectively a time-varying function depending on the active clients set $\mathcal{S}^{(t)}$ as shown in Algorithm 1 (line 11). The convergence and consistency in practical scenarios remain an open problem.

Recently, Wang et al. [261] further showed that the weighting scheme is influenced by the computational heterogeneity across clients as well. If we still use the original weighting scheme, then When clients perform different number of local steps in vanilla FEDAVG and many other FL algorithms, the algorithm will implicitly assign higher weights to the clients with more local steps. Wang et al. [261] give an analytical expression of the surrogate objective function which the original FEDAVG actually optimizes. The authors propose a simple method FEDNOVA to eliminate the inconsistency between the surrogate loss and the original one by “normalizing” (or dividing) the local

updates with the number of local steps. This method can also be considered as a new weighting scheme, as it equivalently assigns lower weights to the clients with more local steps. Another recent work [263] further generalizes the above analysis by showing that the inconsistency problem can appear even with the same number of local steps when using adaptive methods (*e.g.*, ADAGRAD, ADAM) on clients.

We should mention that while the objective inconsistency problem can greatly hurt the performance of FL algorithms even in some simple quadratic models [261], its influence may differ across practical training tasks, depending on the properties of the datasets and neural network models. Due to the non-convex nature of neural networks, it is possible that the minimizer of the surrogate objective is also one of the original objective. On the other hand, if the convergence of the original global objective must be guaranteed in certain applications, then one may carefully choose the weighting scheme according to the suggestions above.

Beyond weighted averaging – neuron matching algorithms In FEDAVG, parameters of local models are averaged coordinate-wise with weights proportional to sizes of the client datasets [181]. One potential shortcoming of FEDAVG is that coordinate-wise averaging of weights may have drastic detrimental effects on the performance of the averaged model and adds significantly to the communication burden. This issue arises due to the permutation invariance of the hidden layers in a neural network. That is, for most neural networks, one can form many functionally equivalent networks (*i.e.*, each input is mapped to the same output) simply by permuting neurons within each layer. Unfortunately, applying coordinate-wise averaging to such functionally equivalent networks may result in a network that produces drastically different output values.

Several recent techniques address this problem by matching the neurons of client NNs before averaging them. For example, Probabilistic Federated Neural Matching (PFNM) [294] utilizes Bayesian non-parametric methods to adjust global model size according to the heterogeneity in the data. In many cases, PFNM has better empirical performance and communication efficiency than FEDAVG. Unfortunately, structural limitations of the method restrict it to relatively simple neural architectures, such as fully connected neural networks with limited depth [257].

Federated Matched Averaging (FedMA) [257] and Model Fusion [229] extend PFNM to other architectures, including CNNs and LSTMs. To address the reduced performance of PFNM on deeper networks, FedMA conducts a layer-wise neural matching scheme. First, the server gathers only the weights of the first layers from the clients and performs one-layer matching to obtain the first layer weights of the federated model. The server then broadcasts these weights to the clients, which proceed to train all consecutive layers on their datasets, keeping the matched federated layers frozen. This procedure is then repeated up to the last layer for which we conduct a weighted averaging based on the class proportions of data points per client. Empirically, FedMA has the potential to achieve better communication efficiency compared to FEDAVG on several FL tasks [257]. In the same vein as FedMA and PFNM, [229] propose the use of optimal transport to match the client neurons before averaging them.

4 On Evaluating Federated Optimization Algorithms

In this section, we discuss how to perform effective evaluations of federated optimization algorithms. Our focus is on topics that are either not present in centralized algorithm evaluation, or exacerbated by facets of federated learning. In particular, we focus on *simulated* federated learning settings, not actual deployments of federated learning. For discussion of production federated learning systems, see Section 5. In this section, we aim to provide examples, discussions, and recommendations focused on federated learning research, though they may also be useful for the practitioner alike.

4.1 Example Evaluations

In order to facilitate this discussion and make it concrete, we implement our recommendations in representative simulated evaluations of three federated optimization algorithms across four datasets.

Federated optimization methods We perform our empirical studies using three existing federated optimization algorithms. We refer to these as Algorithm A, Algorithm B, and Algorithm C. The interested reader can find the details of these algorithms in Appendix B. We omit their names throughout this section since we are not interested in determining which algorithm outperforms the others. Instead, we are interested in how to effectively evaluate the methods. For the purposes of this section, the salient details are that all three algorithms are a special case of Algorithm 1, and employ the same amount of communication and client computation per round. Algorithms A, B, and C all have a client learning rate η and server learning rate η_s , which we tune via grid-search. Instead of performing τ_i steps of CLIENTOPT as in Algorithm 1, we perform E epochs of training over each client’s dataset. This is due to the fact that in practical settings, clients may have local datasets of varying size. Thus, specifying a fixed number of steps can cause some clients to repeatedly train on the same examples, and cause others to only see a small fraction of their data. This epoch-based approach was used in the first work on FEDAVG McMahan et al. [181]. Unless specified, in a given figure, we set $E = 1$.

Datasets We use four federated datasets: GLD-23k, GLD-160k [19, 122], CIFAR-10 [152], and Stack Overflow [18]. The Landmark datasets (GLD-23k and GLD-160k) and Stack Overflow dataset have naturally-arising federated structures. In the Landmark datasets, clients correspond to individual Wikipedia image contributors (for details, see the 2019 Landmark Recognition Challenge [269]). In the Stack Overflow dataset, clients correspond to users of the Stack Overflow forum. For CIFAR-10, we create a non-IID, synthetic client structure by partitioning CIFAR-10 among 10 clients using a form of Dirichlet allocation [121]. The partitioning among clients is identical to that provided by TensorFlow Federated [248], which we mention because consistent partitioning is vital to deriving fair comparisons between methods. See Appendix A for more details on all datasets.

Models and tasks For the Landmark datasets, we predict the image labels using a version of MobileNetV2 [222], replacing the batch norm layers with group norm [273], as suggested in the context of federated learning by Hsieh et al. [120]. For CIFAR-10, we predict the image labels using ResNet-18 (again replacing the batch norm layers with group norm). For Stack Overflow, we train a moderately-sized transformer model to perform next-word-prediction.

Client sampling For the Landmark datasets and Stack Overflow dataset, we only sample a fraction of the training clients at each round, thus simulating cross-device FL. Specifically, clients are sampled uniformly at random, choosing M clients without replacement in a given round, but with replacement across rounds. For CIFAR-10, we sample all 10 clients at each round, simulating cross-silo FL.

In our simulations, we vary the client learning rate, server learning rate, and number of training epochs per client. For more details on the algorithms, models, hyperparameters, and implementations, see Appendix B. We use only one set of experimental setting (dataset and model) to illustrate the general guidelines in Section 4.2. More results on other experimental settings can be found in Appendix C. The code to reproduce all the experimental results is open-sourced².

²https://github.com/google-research/federated/tree/aa2df1c7f513584532cac9c939ddd94f434ed430/fedopt_guide

4.2 Suggestions for Evaluations

This section contains advice for researchers and practitioners on how to evaluate and compare federated optimization algorithms. We stress that our recommendations are neither comprehensive nor always necessary. Rather, we hope they will inspire the reader to think critically about empirical studies of federated optimization algorithms, especially aspects that might differ from empirical evaluations of centralized optimization algorithms. We hope this can be useful in designing new empirical studies, as well as discerning the merits of past and future empirical work on federated optimization.

4.2.1 Use Realistic Tuning Strategies

While centralized learning problems often have explicit training/validation/test splits, this may not be the case in federated learning problems. In order to perform effective tuning for the purposes of simulations, one may wish to use some form of held-out validation data. Notably, such held-out data can be formed in a variety of ways in federated learning.

One simple approach to creating held-out data is to use a held-out set of clients for evaluation. These clients will not participate in any training round. This approach is generally amenable to cross-device settings, where there may be more clients than can reasonably be communicated with during the entire training procedure, and the goal is often to train a model that generalizes to clients not seen during training. In fact, if the number of potential training clients is sufficiently large, it may be statistically valid to simply validate across a random set of training clients, as this validation set will have little overlap with the set of training clients. By contrast, this approach may not be suitable in cross-silo settings, where we have a small set of clients and close to full participation per training round.

Another strategy for generating held-out data that is suitable in cross-silo settings would be to reserve some fraction of the data on each client for validation. This approach has the advantage of ensuring that the validation set is representative, but requires each client to have enough data to effectively partition between train and validation sets. Thus, this approach may not be advisable in cross-device settings where clients have limited amounts of data.

We stress that while the approaches above can be effective for simulations and production use-cases, held-out set tuning may not be the ideal way to perform federated hyperparameter tuning. For example, system constraints (such as limited client availability) may make it difficult to re-run an algorithm with many different hyperparameter settings. Thus, it may be beneficial to instead learn hyperparameters in tandem with the model. Such an approach was used by Khodak et al. [144], who found that hyperparameters can be efficiently learned across clients using techniques based on weight-sharing for neural architecture search [200]. Developing effective federated hyperparameter tuning techniques is an important, but relatively open problem.

Many previous works in federated optimization (including the original paper on FEDAVG [181]) tune their methods on a held-out centralized test set. This can be useful when trying to understand the fundamental limits of federated learning compared to centralized learning. This “test-set tuning” is not usually a practical method, but instead can be viewed as addressing questions like “*In an ideal scenario, can federated optimization methods recover the same accuracy as centralized optimization methods?*”.

Example To exemplify one of the tuning strategies discussed above, we give an explicit example on the Stack Overflow task. This dataset partitions clients into explicit train, validation, and test sets. We compare Algorithms A, B, and C on this dataset, tuning the client and server learning rates by selecting the values maximizing the accuracy of the model on the validation set (see Figure 2).

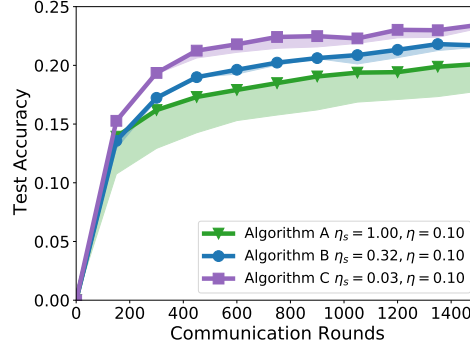


Figure 1: Test accuracy on Stack Overflow for Algorithms A, B, and C. The best and second-best performing client and server learning rate combinations are selected for each algorithm based on validation set accuracy (see Figure 2). The dark lines indicate the test set accuracy for the best-performing combination (given in the legend), with shading indicating the gap between the best and second-best performing runs. This shaded gap provides an indication of the relatively sensitivity of the different algorithms to the resolution of the hyperparameter tuning grid; here we see that Algorithms B and C both perform better and may be easier to tune.

To give a sense for how robust the algorithms are to this tuning, we also determine the second-best configuration of learning rates. The test accuracy for these configurations is given in Figure 1.

4.2.2 Tune Client and Server Learning Rates

The generalized FEDAVG formulation provided in Algorithm 1 introduces the notion of separate learning rates for the client gradient updates and the server update. Learning rate tuning is important both for finding optimal operating points of a new algorithm, and for generating fair comparisons between algorithms. In particular, a server learning rate of 1.0 is not always optimal, even for FEDAVG [211]. Thus, we strongly recommend tuning both learning rates.

Furthermore, because client and server learning rates can be interdependent in some algorithm designs, these should generally be tuned *simultaneously*. This is a key point: to minimize the number of parameter combinations tested, one may wish to tune them serially, perhaps by fixing one, then tuning the other. This can potentially (but not always) lead to sub-optimal operating points.

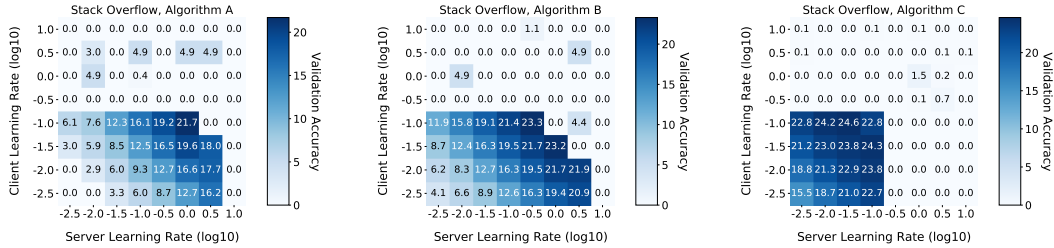


Figure 2: Validation accuracy on Stack Overflow at the last training round (2000), for various client and server learning rates. Results for Algorithms A, B, and C are given in the left, middle, and right plots, respectively. The best and second-best (η_s, η) combinations are used in Figure 1.

For example, in Figure 2, we plot the test accuracy for Algorithms A, B, and C on Stack Overflow as a function of both client and server learning rate. This gives us a full picture of the best possible results achievable by the algorithms with various hyperparameter settings. For Algorithm A, there is only one combination that obtains over 20% accuracy. For Algorithm B, there are multiple configurations that achieve over 20%. However, the client and server learning rates appear to be inter-dependent, and must be set inversely proportional to one another in a specific manner. Thus, algorithms A and B may require tuning client and server learning rates *simultaneously* in order to find near-optimal operating points. At the very least, the tuning must account for the “staircase” shape of the tuning grid.

By contrast, we see that for Algorithm C, 20% accuracy is obtained for all points in our grid where either the client learning rate or server learning rate is 0.1 (as long as the other is at most 0.1). Informally, the rectangular shape of the tuning grid suggests that for Algorithm C, we can first tune the server learning rate, then the client learning rate (or vice-versa). This avoids a quadratic blow-up in the amount of hyperparameter tuning needed.

4.2.3 Analyze Communication-Limited Performance

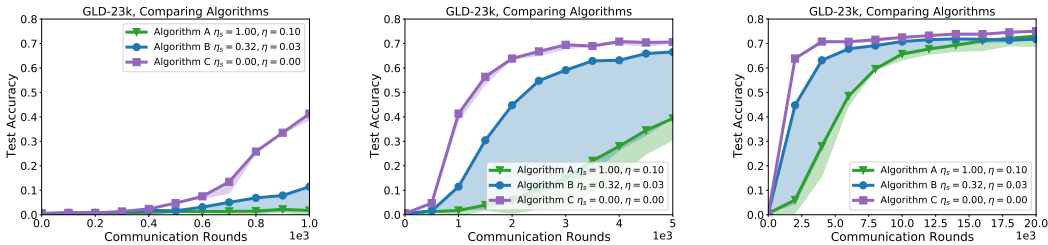


Figure 3: Test accuracy on GLD-23k for a total of 1000, 5000, and 20,000 communication rounds (left, middle, and right, respectively). The best and second-best performing client and server learning rate combinations are selected for each algorithm based on the test set accuracy after the final communication round. The dark lines indicate the test set accuracy for the best-performing combination (given in the legend), with shading indicating the gap between the best and second-best performing runs.

In optimization theory, convergence rates are often asymptotic. Even in empirical evaluations, many methods are evaluated on their performance after sufficiently many training rounds. Given that communication is a fundamental bottleneck in practical federated learning, we encourage authors to instead focus on algorithmic performance in communication-limited settings, especially for cross-device applications. We stress that the relative performance of various algorithms may change depending on the number of communication rounds performed.

For example, in Figure 3, we plot Algorithms A, B, and C for various regimes of communication rounds on the GLD-23k task. Depending on the total number of communication rounds, the conclusions one might draw are different. In the left plot, we might surmise that Algorithms A and B make little progress, while algorithm C is much better. In the middle plot, we might assert that Algorithm A is worse than B and C, and that C is better than B. In the right plot, we see that all algorithms perform equivalently after sufficiently many rounds (as is often required for theoretical results), with B and C achieving slightly better initial accuracy. Notably, Algorithms A and C have little difference between the best and second best learning rate configurations in the right-hand plot, unlike Algorithm B. In short, the actual desired number of communication rounds impacts the conclusions that one might draw about an algorithm. We encourage researchers to be up front about

the success and limitations of their algorithms as the amount of communication changes.

4.2.4 Treat Local Training Steps as a Hyperparameter

According to McMahan et al. [181], one of the key advantages of using FEDAVG over FEDSGD is improved communication-efficiency. This is achieved by having clients perform multiple training steps during each communication round (as opposed to FEDSGD, in which clients effectively perform a single training step). Intuitively, larger numbers of local training steps (τ_i in Algorithm 1) will lead to a reduction in the total number of communication rounds needed to converge. However, as shown repeatedly in the literature on federated optimization, larger numbers of local training steps can also reduce the quality of the learned model [43, 44, 161, 181]. Thus, this strategy often requires appropriate setting of the number of local steps. To obtain the best trade-off between convergence and accuracy, it is beneficial to rigorously tune this hyperparameter.

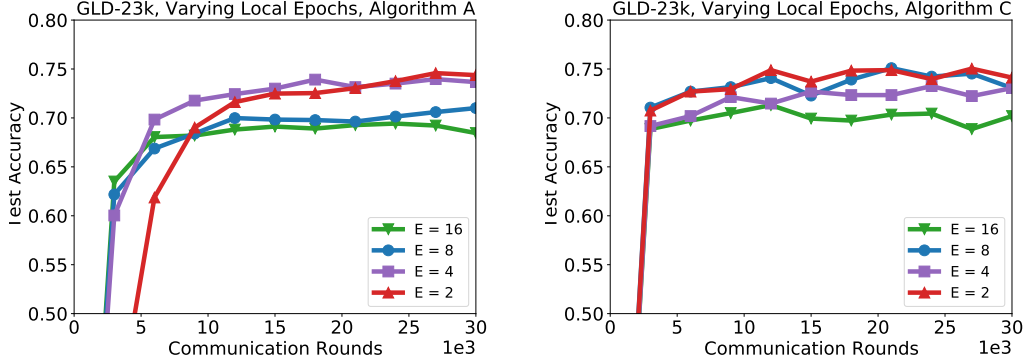


Figure 4: Test accuracy on GLD-23k for Algorithms A (left) and C (right) for various numbers of local epochs per round E , versus the number of communication rounds. We set $\eta = 0.1, \eta_s = 1.0$ for Algorithm A, and $\eta = 0.01, \eta_s = 10^{-5/2}$ for Algorithm C.

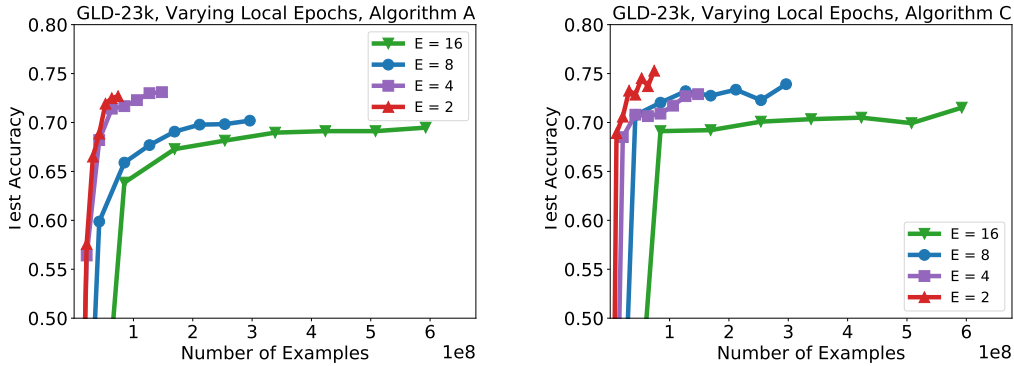


Figure 5: Test accuracy on GLD-23k for Algorithms A (left) and C (right) for various numbers of local epochs per round, versus the total number of examples processed by all clients. We set $\eta = 0.1, \eta_s = 1.0$ for Algorithm A, and $\eta = 0.01, \eta_s = 10^{-5/2}$ for Algorithm C. These learning rates were chosen as they performed well for $E = 2$. While they also performed near-optimally for other values of E , tuning the learning rates jointly with E may produce slightly better results.

For example, in Figure 4, we apply Algorithms A and C with different number of local epochs E on the GLD-23k task. We see that for Algorithm A, the choice of number of local epochs can make a large impact on the initial accuracy of the model, and minor differences in later rounds. On the other hand, Algorithm C is robust against the choice of E , achieving comparable accuracy for all settings used throughout.

While Figure 4 is useful for judging the number of communication rounds needed, it omits the amount of *total* computation being performed by the clients. For each client, $E = 16$ involves $8\times$ as much work as $E = 2$, something that is particularly important in cross-device settings where clients may have limited compute capacity. To judge this overall computation complexity, we give an alternate view of these results in Figure 5. Instead of plotting the number of communication rounds, we plot the total number of examples processed by all clients throughout the training process. We see that this gives a drastically different view of the algorithms and the role of E . We see that for both algorithms, $E = 16$ actually processes many more examples than $E = 2$ to reach a comparable accuracy. Thus, benefits of increasing E saturate, and even decrease. However, we see that Algorithm C is slightly more robust to settings of E , achieving roughly the same “example-efficiency” for $E = 2, 4, 8$, while Algorithm A only sees comparable accuracy for $E = 2, 4$.

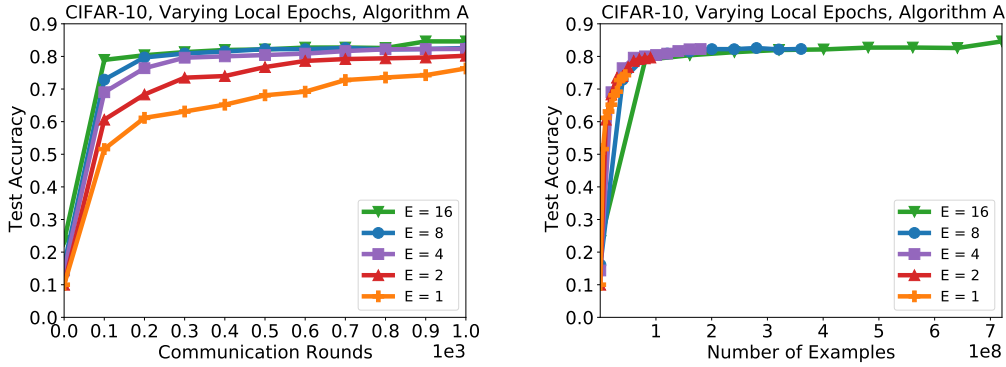


Figure 6: Test accuracy on CIFAR-10 for Algorithms A. We plot test accuracy versus the number of communication rounds (left) and versus the total number of examples processed by clients (right) for various numbers of local epochs per round E . We use learning rates $\eta = 0.01, \eta_s = 1.0$. These learning rates were chosen as they performed well for $E = 2$. While they also performed near-optimally for other values of E , tuning the learning rates jointly with E may produce slightly better results.

We also note that such findings may depend on the federated setting being simulated. For example, we perform the same experiment as above with Algorithm A, but for CIFAR-10. The results are in Figure 6. While we see similar results to the left-hand plots of Figures 4 and 5, the plots are slightly different when comparing accuracy versus number of examples seen. In particular, while Algorithm A can obtain higher accuracies with fewer examples when using a smaller value of E on GLD-23k, we see roughly comparable numbers of examples processed for any test accuracy on CIFAR-10. In other words, Algorithm A experiences diminishing returns (in terms of data efficiency) on GLD-23k. It does not see such diminishing returns on CIFAR-10. This may be due to the difference in nature of the two datasets. GLD-23k is a more heterogeneous, cross-device dataset, and therefore setting E to be too large can be detrimental [270], while CIFAR-10 is more homogeneous and cross-silo, and therefore may benefit from larger E [271].

4.2.5 Understand the Impact of Cohort Size

One fundamental aspect of (cross-device) federated optimization that we have not yet broached is the effect of the number of clients M sampled at each round (i.e., the cohort size). Intuitively, the larger M is, the less variance in the aggregate client updates. This in turn could lead to a reduction in the total number of training rounds needed to obtain a given accuracy. For example, in Figure 7, we plot the accuracy of Algorithms A and C on GLD-160k, varying the cohort size M over $\{10, 40, 80\}$. We see that for both algorithms, increasing M leads to a better final accuracy. For Algorithm C, we also see a dramatic increase in the initial accuracy of the learned model.

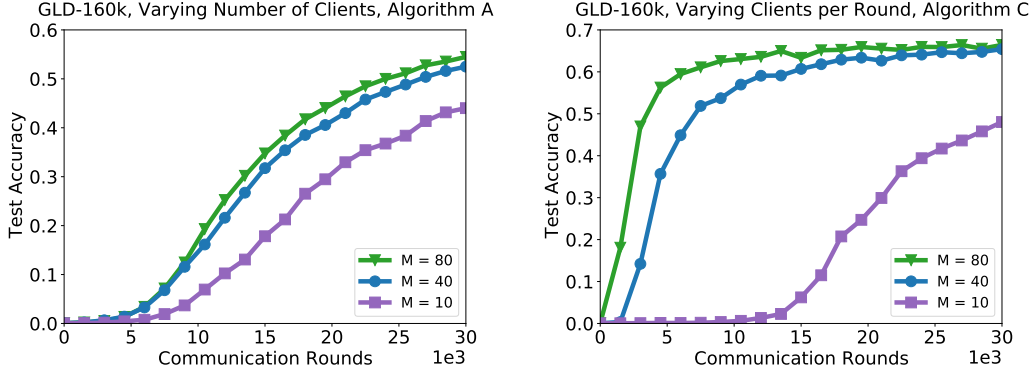


Figure 7: Test accuracy on GLD-160k for Algorithms A (left) and C (right) for various cohort size M , versus the number of communication rounds. We use learning rates $\eta = 0.1, \eta_s = 1$ for Algorithm A, and $\eta = 0.01, \eta_s = 10^{-5/2}$ for Algorithm C. These learning rates were chosen as they performed well for $M = 10$. While they also performed near-optimally for other values of M , tuning the learning rates jointly with M may produce slightly better results.

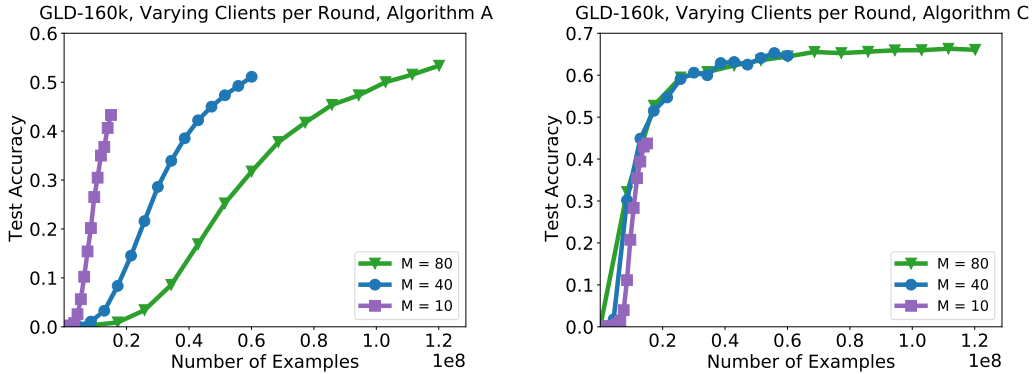


Figure 8: Test accuracy on GLD-160k for Algorithms A (left) and C (right) for various cohort sizes M , versus the number of examples processed by the clients. We use learning rates $\eta = 10^{-3/2}, \eta_s = 1$ for Algorithm A, and $\eta = \eta_s = 10^{-5/2}$ for Algorithm C.

Just as in Section 4.2.4, comparing these settings with respect to the number of communication rounds omits some key information. While all clients are now performing the same amount of computation per round, the total amount of client computation per round varies depending on M . Increased total client computation time can have negative impacts, both due to electricity usage

and carbon emissions, as well as delayed round completion times stemming from straggler effects. In Figure 8, we give an alternative view of this experiment, plotting test accuracy versus the total number of examples processed at all clients. We see that the story changes dramatically for Algorithm A and C. Algorithm C seems to achieve comparable accuracy for a given number of examples, for all settings of M . For Algorithm A, higher values of M actually require many more examples to be processed to reach smaller accuracy thresholds. This is an important point: The benefits and drawbacks of increasing M can vary significantly depending on the federated optimization algorithm employed. More experimental results and discussion on how the number of clients sampled per round can affect federated training dynamics can be found in [45].

4.2.6 Other Considerations

Motivate simulation designs with well-specified target applications The guidelines in Section 4.2 are generally applicable to both cross-device and cross-silo settings. While the example algorithms A, B, C were initially designed for cross-device federated learning, they are also applicable to cross-silo federated learning. However, as discussed in Section 1.2, cross-device and cross-silo settings have important distinctions. Echoing the suggestion in Section 3.1.1 for algorithm design, we encourage researchers to clearly specify the application setting of any evaluations. For example, when evaluating a method in cross-silo settings, it is not necessary for it to scale to millions of clients. By contrast, evaluating scalability of an algorithm may be vital when considering cross-device settings. While methods suitable for cross-device settings can often apply to cross-silo settings without significant modifications, the reverse is generally not true. More sophisticated cross-silo algorithms may require global information across clients, or require stateful clients (for example, SCAFFOLD [138]). Thus, it is important to be clear about what settings are being considered in any evaluation of federated optimization.

Further, it will usually not be realistic to explore all possible combinations of local epochs E , cohort sizes M , client and server learning rates, rounds of communication, etc. Thus, when decisions need to be made which limit the scope of experiments, it is best to make choices based on what is likely to be practical and realistic for the targeted application; comparing to the choices made in prior works can also be valuable, but if those regimes are not relevant to the target application, then such comparisons can be optional.

Using alternate evaluation metrics In the examples above, we report the accuracy of a model over all test samples (*i.e.*, the accuracy is the number of correct predictions over the entire test set, divided by the number of test samples). This ignores potential data imbalances across test clients. To perform better evaluations of federated algorithms, one may wish to use alternate evaluation metrics. For example, one could look at the mean test accuracy across all test clients. More generally, one could look at the distribution of a model’s accuracy across clients. This is particularly important in areas such as fairness, where quantities such as the minimum accuracy across all clients play an important role (see Section 7.3 for more discussion on this topic). Developing better evaluation metrics for federated learning is an open problem, and we encourage researchers to clearly specify the metrics used for experimental evaluation.

Alternate metrics aggregations are particularly useful when trying to measure the fairness of a model. For example, one measure of fairness is the minimum accuracy across all clients. Of course, there are a number of other ways this could be examined, and we caution the reader that there is no single way of computing the fairness of a model (see work by Li et al. [162] and Hu et al. [123] for various approaches to measuring fairness in empirical evaluations). We encourage authors to be explicit about the measure they are using, and its potential drawbacks. For example, the minimum accuracy measure above guarantees a floor on client accuracy, but does not measure the

variance of accuracy across clients (such as the difference between the maximum and minimum client accuracy). Developing useful fairness metrics is an ongoing and important area of study within federated learning.

Isolate sources of randomness Empirical evaluations are typically stronger when they try to account for any sources of randomness (e.g., model initialization). Federated learning has notable sources of extra randomness. In cross-device settings, which clients are sampled at which round has the potential to significantly affect the learned model (particularly when M is small), in much the same way that the order of examples seen while performing SGD (especially non-random orders) can alter the underlying training dynamics [75]. In order to derive fair comparisons, we recommend using pseudo-random client sampling in simulations to see how algorithms behave with the same sequence of sampled clients (see [211] for an example of doing this in federated settings). More generally, we recommend varying the random seed governing the pseudo-random sampling and plotting the mean and variance of results across trials. More sophisticated empirical analyses may even perform biased sampling of clients in order to simulate things like client availability and heterogeneity in device capabilities [75].

Use learning rate scheduling One common technique used to train centralized models is learning rate scheduling, in particular warmup and decay [98]. Such techniques can be vital for training large models such as transformers and deep ResNet models. Thus, learning rate scheduling may also be useful in federated learning. In fact, Li et al. [166] show that some form of learning rate decay is *necessary* to guarantee that FEDAVG reaches critical points of the empirical loss function. How exactly to configure such scheduling is an open question. For example, in Algorithm 1, learning rate scheduling can be done on the client or on the server (or both). If clients perform many local training steps, then client learning rate decay may be effective. However, if clients only perform a small number of updates relative to the number of communication rounds, server learning rate decay may be a more effective strategy.

4.3 On the Role of Toy Problems and Artificial Datasets

Due to the nascent and sometimes counter-intuitive findings of federated optimization, synthetic datasets can be useful for performing an initial validation of a method. Such datasets can be particularly useful for understanding federated optimization algorithms in controlled environments. For example, one might use such datasets to isolate and analyze the effect of single factor. Such datasets are often most useful as a way to gain insight over a single facet of federated learning, especially in service of subsequent theoretical analyses. Below we discuss some examples of toy datasets, their utility, and limits on their ability to provide meaningful simulations of federated learning.

4.3.1 Quadratic and Convex Problems on Synthetic Data

A frequently used class of toy problems are quadratic problems, *i.e.*, problems where the client objective functions $F_i(\mathbf{x})$ in (1) are quadratic functions. For instance, Charles and Konečný [44] consider functions of the form $F_i(\mathbf{x}) = \frac{1}{2} \|A_i^{1/2}(\mathbf{x} - \mathbf{c}_i)\|^2$ for $A_i \in \mathbb{R}^{d \times d}$ symmetric and positive semi-definite and $\mathbf{c}_i \in \mathbb{R}^d$. Such formulations are useful for understanding the impact of heterogeneity among client datasets, as one can reduce the heterogeneity of the datasets to the heterogeneity of the pairs (A_i, \mathbf{c}_i) . The quadratic nature also frequently allows for closed-form expressions of quantities of interest. For example, for the quadratic functions above, each client’s optimal model is given by

$\mathbf{x}_i^* := A_i^\dagger A_i \mathbf{c}_i$ (where A^\dagger denotes the pseudo-inverse of a matrix A). These optimal points for each client are in general different from the globally optimal solution, $\mathbf{x}^* = (\mathbb{E}_{i \sim \mathcal{P}} A_i)^\dagger (\mathbb{E}_{i \sim \mathcal{P}} A_i \mathbf{c}_i)$.

Another approach to generating a synthetic quadratic problem was formulated by Li et al. [161] using heterogeneous Gaussian distributions. In their approach, each client i has samples x drawn from a Gaussian $\mathcal{N}(\mu_i, \Sigma)$. Each x has label given by a generalized linear model $y = \sigma(W_i x + b_i)$. Notably, μ_i, W_i, b_i are all random variables. While the exact details of the distributions can vary, this toy problem allows one to control the heterogeneity of the μ_i (*i.e.*, the heterogeneity of the clients' examples) separately from the heterogeneity of the w_i, b_i (*i.e.*, the heterogeneity of the clients' generalized linear models).

Thematically similar quadratic and convex toy problems have been used in various works on distributed optimization [147, 161, 226, 261]. While frequently useful, especially from a theoretical point of view, there are a number of pitfalls to be cognizant of when using them. For example, quadratic functions are not a representative benchmark for balanced, IID, datasets due to their functional simplicity [271]. More generally, convex problems are not representative benchmarks of non-convex problems, due to their inherent property that $F(\mathbb{E}_{i \sim \mathcal{P}} \mathbf{x}_i^*) \leq \mathbb{E}_{i \sim \mathcal{P}} F(\mathbf{x}_i^*)$, *i.e.*, the average of the clients' models over-performs. This property does not transfer to non-convex tasks.

4.3.2 Artificial Partitioning of Centralized Datasets

Another common way to create synthetic federated datasets is to take a centralized dataset, and employ some scheme to partition the dataset among a set of clients. Compared to the synthetic data approach above, this has the advantage of allowing one to use more realistic examples and well-understood models. One particularly prevalent approach is to take a labeled dataset (such as CIFAR-10) and partition examples by randomly distributing the set of all examples with label ℓ across some set of clients S_ℓ . While this approach does result in a heterogeneous dataset, it is not necessarily indicative of realistic heterogeneous datasets. Thus, we caution the reader that such pathological partitioning should be done primarily for the purpose of understanding specific and extreme forms of heterogeneity. In the example of label partitioning above, this approach is most useful when trying to specifically understand the impact of label heterogeneity (*i.e.*, when each client may only have a small subset of the total set of labels) on a federated algorithm.

Recent work has attempted to partition centralized datasets using more realistic, but controllable partitioning schemes. Notably, Hsu et al. [121] propose a useful partitioning scheme based on Latent Dirichlet Allocation from topic modeling. This has the advantage of allowing one to control the amount of heterogeneity by changing the parameter α of the underlying distribution. In fact, varying α in $[0, \infty)$ allows the user to interpolate between a completely IID partitioning of examples and the pathological label partitioning scheme above (wherein each client's dataset has a single label). Later work by Reddi et al. [211] used the Pachinko Allocation Method [165] to partition datasets with a hierarchical label structure (such as CIFAR-100).

While these topic modelling approaches can generate more realistic artificial federated datasets, we caution that they are not necessarily emblematic of realistic federated datasets. For example, such methods assume that all clients generate data from the same stochastic process, which may not reflect datasets with dissimilar users. Instead, we argue that these topic modelling approaches are most useful when trying to understand the impact of heterogeneity on an algorithm. Because the underlying topics are generally controllable via some user-specified distribution, the user can evaluate an algorithm for various amounts of heterogeneity in order to perform effective ablation studies (see [121] for valuable examples of this).

4.3.3 Datasets with Examples Shuffled Across Users

As discussed in Section 1.1, a challenge for many federated algorithms is the presence of heterogeneous data across clients. Notably, this also poses a challenge for performing effective ablation studies on a federated optimization algorithm. Importantly, the effect of heterogeneous data across clients can be effectively removed by amalgamating all client datasets, and creating a new dataset where each client’s dataset is drawn uniformly at random from this super-set of examples. More sophisticated ablation studies could also interpolate between this “shuffled”, homogeneous dataset and the original dataset, for instance by having each client use a kind of mixture between their original dataset, and some global pool of examples.

While similar to standard server evaluation, this type of evaluation allows one to test novel federated algorithms in a simplified setting. This can be useful as a first step towards determining the plausibility of an idea, or as a way to measure the effect of client drift on an algorithm (eg. by evaluating an algorithm on both the original and shuffled dataset). However, such shuffled datasets are not indicative of the performance of federated optimization algorithms in practical settings. We therefore recommend such “pathological” datasets as ways to perform more fine-grained evaluation, not as a way to actually benchmark federated optimization methods.

5 System Constraints and Practices

In Section 3 we discussed methods to improve algorithm performance by considering heterogeneity, communication efficiency and many other factors; and in Section 4 we evaluated algorithms in the absence of the deployment system. In this section, we will examine the dimensions of heterogeneity and other factors that appear in practice and discuss evaluating federated algorithms on real-world federated learning systems such as Bonawitz et al. [39], Paulik et al. [199]. We examine how the costs of communication, computation, and memory differ from the simulations used for developing algorithms, and what effect that has on real-world deployment. Then we will propose a basic model parameterized by system properties that can be used to evaluate algorithm deployment costs. We aim to show that co-designing numerical optimization and system optimization can achieve practical federated learning systems and applications. Finally, we will conclude with suggestions for deploying algorithms on real-world systems.

5.1 Communication Costs

When developing and experimenting with federated learning algorithms, researchers and engineers typically use simulation environments. In simulated environments, communication can be effectively free, and typically always available, as the computation is often co-located either on a single machine or within a highly interconnected data center. This setup does not reflect real-world federated learning systems. In cross-device settings, compute power and communication capacity can exhibit extreme heterogeneity. Even in cross-silo settings, where compute resources may be less constrained, participants in a federated learning algorithm may be geographically separated, incurring large communication costs.

When we talk about communication-efficient algorithms in academia, we often measure using the frequency of communication, which is only part of the whole picture. In contrast, in real-world deployments, the elapsed time required to train a model is an important but often overlooked concern. It is important to also consider what factors contribute to where time is spent, as there are multiple dimensions to consider even within communication cost.

Redundancy in model updates As mentioned in Section 3.1, compressed communication can often help in federated learning. Communicated matrices may have many forms of redundancy (*e.g.*, sparsity, spatial), allowing for significant compression ratios. For instance, if the model updates are low rank, adopting low-rank/sketching-based compression methods [149, 253, 255] can achieve high communication efficiency. This may indicate a metric that understands the distribution of expected values, not only the shapes and sizes of the communications, is desirable.

Bandwidth heterogeneity and dynamics Unlike distributed training in the data center setting, where network bandwidths are usually homogeneous and stable over time, in the cross-device setting edge devices typically communicate with a central server via potentially low quality network connections. In real-world deployments, the available bandwidth of wireless networks may have vastly different download versus upload capabilities. Furthermore, wireless networks can be noisy, are prone to interference, and vary in quality over time.

Stragglers In synchronous learning algorithms with fixed cohorts (such as FEDAVG as described in Section 2.1), communication cost of the system can be dominated by the slowest participant, as round completion is blocked by stragglers [39]. Further, it is not always the case that the weakest compute node is also the weakest network link: communication-computation ratios can be highly imbalanced within a round. To mitigate the effect of extreme stragglers both Bonawitz et al. [39] and Paulik et al. [199] describe participant reporting deadlines for synchronization points, and recent research demonstrates algorithms that can mitigate the effect of slow participants [161, 203, 213]. These may not be necessary in cross-silo settings, where dedicated compute hardware and network availability can be more reliable.

Extra cost for privacy and security: secure aggregation It is desirable for FL system to accommodate various privacy and security methods. However, such methods may introduce extra cost in the system. One example is the cryptographic protocol SECAGG [38] used to protect the privacy of the participants during the communication back to the central aggregator. SECAGG relies on shared secret keys for each pair of participating users, which adds quadratic complexity to the beginning of every communication round. Alternatively, [115] requires a second non-colluding auxiliary server, giving a secure scheme with computation and communication overhead of at most $2\times$. To perform single-server SECAGG efficiently, Bell et al. [32] presents a protocol where both client computation and communication depend logarithmically the number of participating clients, potentially enabling scaling the protocol to much greater numbers of clients per communication round. Turbo-Aggregate [234] employs a multi-group circular strategy for efficient model aggregation, and leverages additive secret sharing and coding techniques to handle user dropouts while guaranteeing user privacy, which has a complexity that grows (almost) linearly with the number of the users, while providing robustness to 50% user dropout. Furthermore, encryption generally reduces compression ratios, as it tries to make the data indistinguishable from random values. Compatibility limitations of secure computation with federated algorithms are further discussed in Section 7.1.1.

5.2 Computation and Memory Costs

As described in Section 3.1, performing optimization computations at edge devices, such as local SGD and client drift mitigation, incurs computation cost of the client. In cross-device settings, participants are generally resource-constrained devices such as smartphones. We must take such computation limitations into consideration to design computation-efficient federated optimization strategies that are viable for real-world deployments.

Heterogeneous resources Large variability in training participant resources can lead to difficulty in real-world systems, particularly in cross-device settings. High-end devices with powerful computation and memory resources can train larger models while low-end devices can only train smaller models. Systems that “close” a round of communication once a threshold of participants are finished may unfairly bias results towards those resource-rich participants. How to take advantage of these high-end devices and efficiently aggregate heterogeneous models at the server side remains an open research question.

System policies In cross-device settings where participants are generally smartphone devices, resources are usually shared with other non-learning processes running on the system. System policies may introduce requirements such as the device being in an idle state, connected to particular networks, or charging [39, 199], which can severely limit the devices’ ability to participate in training. Compared to dedicated data center compute resources, this will result in unreliable compute availability and fluctuating resources [79, 285].

Memory constraints Smartphone clients participated in federated optimization have much less available memory than the high performance servers used in centralized data center training and federated learning simulation. The trend in machine learning is ever larger and deeper models. Combined with more complex optimization techniques that require additional parameters (momentum, etc.), training a model quickly becomes infeasible on devices with limited resources. Techniques such as gradient checkpointing [55] can reduce memory usage at the cost of more computation. He et al. [112] investigate performing more computation on the resource-rich server, though such techniques may not apply to fully decentralized federated learning regimes.

5.3 Analytical Communication and Computation Costs

Building a large-scale federated learning system is sometimes infeasible for academic researchers, and inconvenient for exploring and prototyping research ideas. We now discuss approaches to evaluate the practical communication and computation cost before deploying the algorithms in a real system. We’ll examine the scenarios where the communication and computation costs can be analyzed directly, e.g., if two algorithms have identical amounts of computation and communication in one round of training. In this case the conventional metric of “number of total rounds to complete training” is a reliable metric for comparison. This analytical comparison can be used when one algorithm is strictly better on computation and/or communication cost, e.g., fast lossless compression for communication. However, we defer the discussion to Section 5.4 when two algorithms have different computation and communication costs and it is difficult to tell which one is better from simulation only.

In one of the earliest federated optimization papers, McMahan et al. [181] measured the performance of various algorithms in terms of how the training loss varies with the number of communication rounds (which we often refer to as simply a round). Note that this is a useful method for comparing methods with the same amount of communication and local computation. For example, this is used by Li et al. [161] to accurately compare FEDAVG with FEDPROX, and by Reddi et al. [211] to compare FEDAVG with methods such as FEDADAM. However, many works on federated optimization alter the number of local gradient evaluations or the amount of communication per round. In such cases, it is not sufficient to compare the convergence of different methods only in terms of the number of communication rounds. Instead, we suggest using a metric of estimated wall-clock time spent per round in a real-world FL deployment.

In federated learning, wall-clock time can be broken down into the local computation time T_{comp} and the communication time T_{comm} . The local computation time is proportional to the number of gradient evaluations per round, which depends on the optimization algorithm. The

communication time T_{comm} can be formulated as $\alpha_{|\mathcal{S}|} + \beta d$. The parameter $\alpha_{|\mathcal{S}|}$ is a fixed latency for every communication with the central server irrespective of the bits transmitted, which can be an increasing function of the number of active clients $|\mathcal{S}|$. The size of the updates communicated per round is d bits and the resulting latency is βd , where $1/\beta$ is available transmission bandwidth. Observe that the relative values of the fixed latency $\alpha_{|\mathcal{S}|}$ and the message transmission time βd depend on the model size, the optimization algorithm as well as the communication infrastructure underlying the federated learning framework. In Table 1, we provide a few examples of the computation and communication costs in some recent federated optimization algorithms.

Algorithms	Message size	Comm. latency	Grad. evaluations
FEDAVG [181]	$2d$	$\alpha_{ \mathcal{S} }$	τb
FEDADAM/FEDYOGI [211]	$2d$	$\alpha_{ \mathcal{S} }$	τb
FEDPA [6]	$2d$	$\alpha_{ \mathcal{S} }$	τb
SCAFFOLD [138]	$4d$	$\alpha_{ \mathcal{S} }$	τb
MIME/MIMELITE [137]	$4d/3d$	$\alpha_{ \mathcal{S} }$	$\tau b + \mathcal{D}_i $

Table 1: Examples of communication and computation costs in some recent federated optimization algorithms. All metrics are evaluated for one round (*i.e.*, one global iteration). In the table, α is a function describing how the communication latency scales with the number of active clients. The total training time for one round can be formulated as a linear combination of the above three metrics and the weight of each metric varies in different systems. Most previous works use communication rounds as the measure of communication cost. This makes sense when the communication latency dominates the other two terms and the scaling function $\alpha_{|\mathcal{S}|}$ is a constant.

The computation and communication formulation $T_{\text{comp}} + T_{\text{comm}}$ of wall-clock time does not consider the work performed by the central server for global model update. Unlike the edge device participants, the central server is generally not resource constrained in the cross-device setting, and the computation for updating the global model is not significant for generalized FEDAVG algorithms. Next we will propose a system model that enables using measurements of simulation time of federated optimization algorithms to achieve ballpark estimations of real world wall-clock time.

5.4 Basic Model to Estimate Round Time of Cross-Device Training

Motivation We propose a basic system model to estimate the communication efficiency of deploying a federated algorithm to real-world systems. We hope this basic model can serve the following purposes. First, it represents a basic start for systems research to bridge FL simulation and deployment, which is challenging and lacks attention in academia so far. Second, this model can provide a rough estimation to compare the performance of two federated learning algorithms primarily in execution time when deployed in a cross-device federated learning system. Third, this basic model can inspire federated optimization algorithms and system co-design, where federated optimization can be designed and adjusted considering various system constraints for potential practical deployment, and federated systems design can be influenced when considering support for various types of federated optimization algorithms. The model discussed here is simplified, we did not consider stragglers and many other real-world system constraints mentioned in Section 5.1 and Section 5.2. We notice a few recent papers also discussed system modeling in simulation for federated learning [154, 282].

Basic execution model Our basic model will estimate the execution time per round when deploying an optimization algorithm in cross-device federated learning system as follows,

$$\begin{aligned}
T_{\text{round}}(\mathcal{A}) &= T_{\text{comm}}(\mathcal{A}) + T_{\text{comp}}(\mathcal{A}), \quad T_{\text{comm}}(\mathcal{A}) = \frac{S_{\text{down}}(\mathcal{A})}{B_{\text{down}}} + \frac{S_{\text{up}}(\mathcal{A})}{B_{\text{up}}}, \\
T_{\text{comp}}(\mathcal{A}) &= \max_{j \in \mathcal{D}_{\text{round}}} T_{\text{client}}^j + T_{\text{server}}(\mathcal{A}), \quad T_{\text{client}}^j(\mathcal{A}) = R_{\text{comp}} T_{\text{sim}}^j(\mathcal{A}) + C_{\text{comp}},
\end{aligned} \tag{9}$$

where client download size $S_{\text{down}}(\mathcal{A})$, upload size $S_{\text{up}}(\mathcal{A})$, server computation time $T_{\text{server}}(\mathcal{A})$ and client computation time $T_{\text{sim}}^j(\mathcal{A})$ depend on model and algorithm \mathcal{A} . Simulation time $T_{\text{server}}(\mathcal{A})$ and $T_{\text{sim}}^j(\mathcal{A})$ can be estimated from FL simulation in the data center ³; download bandwidth B_{down} , upload bandwidth B_{up} , on-device to data center computational ratio R_{comp} , and system constant C_{comp} are parameters of this system model.

We estimate parameters (B_{down} , B_{up} , R_{comp} , C_{comp}) of this model based on a real world cross-device federated learning system [39], which supports Google products like Gboard [104]. We pick several production tasks in [39], and collect the average metrics for anonymous users participating in these tasks in a limited time window. These tasks involve various models like RNN language models and MLP encoders. For download bandwidth B_{down} and upload bandwidth B_{up} , we estimate the parameters based on the slowest device when 90% of the devices can complete the round and report back to the server for each task, and take the average over different tasks. For computational ratio R_{comp} and system constant C_{comp} , we first estimate simulation time for each task we picked by executing the same model and algorithm on a desktop CPU. Then a linear model in (9) is fitted based on anonymous aggregated data of on-device training time. Finally, we average the coefficients of the linear model over the tasks picked for R_{comp} and C_{comp} .

The estimated parameters of the basic model are

$$B_{\text{down}} \sim 0.75 \text{ MB/secs}, B_{\text{up}} \sim 0.25 \text{ MB/secs}, R_{\text{comp}} \sim 7, \text{ and } C_{\text{comp}} \sim 10 \text{ secs.} \tag{10}$$

The numbers in equation (10) are only intended to provide some insights of real cross-device federated learning systems, and the scales of these parameters can be more important than the absolute values.

Usage When there is no clear conclusion from applying analytical analysis mentioned in Section 5.3 for simulation experiments, the basic model and estimated parameters from the cross-device FL system [39] can be used to compare the speed of federated optimization algorithms for on-device training. In addition, the system model can be used to identify parameter settings (e.g. B_{down} , R_{comp}) where the preferred candidate optimization algorithm may change. We only offer a rough estimation of parameters in the basic model, and these parameters can have large variance due to system configuration like population size and geographical distribution of devices. For example, if an algorithm can tolerate a higher dropout rate (less clients need to finish training and report back to the server), then the upload and download bandwidth parameter (estimated from the real system) in the basic model can be improved and communication cost can be reduced. An optimization algorithm can be claimed to be better than another one in scenarios defined by a certain range of the basic model parameters.

This basic model has limitations. For example, this model is not designed to account for the effect of stragglers, or behavior of various clients. However the model is sufficient to help discuss the settings when a communication efficient algorithm saves more time than a computation efficient algorithm. This basic model is provided here to foster the discussion of more sophisticated models and system algorithm co-design for federated optimization in the future. A simple extension for future

³We assume the simulation environment is a data center similar to actual server in a cross-device FL system, and ignore the constant for server system setup time. The server simulation time is usually small for the variant of FEDAVG algorithms where the aggregation is (weighted) summation of client updates and server computation is one step of optimizer update based on the aggregated update.

work could use a distribution instead of scalar to model the parameters for download bandwidth, upload bandwidth, computational ratio, and system constant.

Example In Figure 9 we plot the test accuracy of a model on the GLD-23k dataset (see Section 4 for the simulation setup), as a function of the number of communication rounds, and the estimated round completion time $T_{\text{round}}(\mathcal{A})$. We use Algorithm A⁴, and vary the number of local epochs of training E each client performs per round. Based on simulations conducted on a single CPU (in order to better emulate the lightweight capabilities of cross-device clients), we estimate that the model used for GLD-23k requires roughly 0.127 seconds to process a single data example. This allows us to estimate $T_{\text{sim}}^j(\mathcal{A})$ in this case as simply 0.127 multiplied by the number of examples processed by a client. Using the estimated constants B_{down} , B_{up} , R_{comp} , and C_{comp} above, we can then estimate $T_{\text{round}}(\mathcal{A})$ using (9). While large local epoch E settings converge faster if we look at the curves for the first 2000 communication rounds, the extra computation at the clients consumes extra training time such that the entire training time slows down drastically. It is worth mentioning that while more rounds can be accomplished under given wall-clock time constraint when E is small, this can have potential negative consequences when combining federated optimization with differential privacy techniques. More results on Algorithm C and the Stack Overflow experiment can be found in Appendix C.

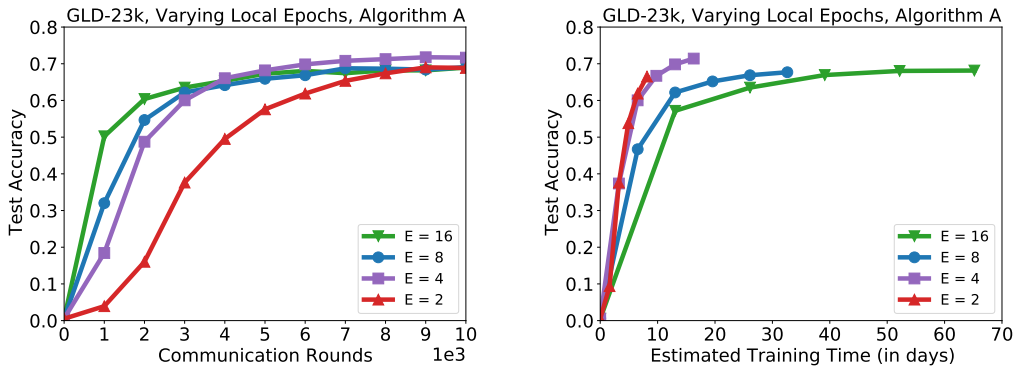


Figure 9: Test accuracy on GLD-23k for Algorithm A, for varying numbers of local epochs E per round. We plot the test accuracy versus the number of communication rounds (left) and an estimate of the execution model round completion time $T_{\text{round}}(\mathcal{A})$ (right). We use learning rates $\eta = 0.1$ and $\eta_s = 1.0$.

5.5 Research Frameworks for Simulating Distributed Systems

In the last few years a proliferation of frameworks have been developed to enable federated optimization researchers to better explore and study algorithms, both from academia (LEAF [41], FedML [113], Flower [34]) and from industry (PySyft [219], TensorFlow Federated (TFF) [125], FATE [284], Clara [191], PaddleFL [175], Open FL [127], FedJAX [216]). In this section we highlight some desirable aspects from such frameworks.

Traditionally many frameworks largely focused on enabling empirical studies of algorithms in a data center for experimental simulation. The community can benefit from frameworks that also facilitate simulating and analyzing the systems aspects such as elapsed training time, communication costs, and computation costs under different architectures.

⁴See the beginning of Section 4 for a discussion on why we obfuscate the algorithm.

A framework for designing federated algorithms that can be analyzed and deployed on real world systems might include features such as:

- *Platform agnostic*: Federated scenarios generally are characterized by heterogeneous compute nodes. The ability to write once and run everywhere is important to reach the participants of the algorithm.
- *Expressive communication primitives*: A flexible programming interface that enables expression of diverse network topologies, variable information exchange among workers/clients, and various training procedures.
- *Scalability*: Both cross-silo settings and cross-device settings have their own specific scaling needs. For example, cross-silo simulations may require scaling to a small number of data-rich clients, while cross-device simulations may require scaling to a large number of data- and compute-limited clients. A framework that can scale in multiple dimensions adds additional flexibility.
- *Extensibility*: new algorithms are developed that can require new capabilities. Keeping pace with the new frontiers of research prevents a framework from becoming obsolete.

For more detailed challenges and practice in these topics, we refer to the guidance provided by [132, Section 7].

5.6 Real-world Deployment Suggestions

A central challenge in deploying practical federated learning systems is the need for modeling without access to raw centralized data. In a typical centralized training workflow, a modeling engineer is able to quickly test their hypotheses by inspecting data samples, *e.g.*, identifying mis-classified examples. This can lead to improved models and can also surface issues with the model or data pipeline. In federated settings, raw data is located on client devices and direct data access for modeling or debugging may be impossible.

De-risk training in simulation with proxy data When available, proxy data (real or synthetic data that can serve as an approximate substitute for federated data) is useful for testing federated algorithms in simulation. Ideally, a practical federated learning system is built so that taking a working simulation and deploying it to real-world clients is as seamless as possible, to minimize bugs introduced in the process.

To ensure that observations made in simulated federated learning transfer well to production settings, simulated client data must have high fidelity. The most important factor for federated simulations is the clustering of data into pseudo-clients. Production federated data are highly heterogeneous, so insights from federated simulations on homogeneous (IID) client data are not guaranteed to generalize. The distribution of the number of training examples per client should also be matched in simulations, as they affect the optimal configuration of client and server optimizer parameters. Length distributions for sequence-based features should also be matched, as data augmentation techniques may need to be re-tuned. As an example, optimizations made for language modeling on long sentences in simulations are not guaranteed to generalize well to on-device settings with shorter sentence lengths. Typically, improvements to simulation and production federated training are iterative. Insights into production failures lead to simulation improvements, which can be used to optimize better models that eventually generalize better on-device.

Consider alternative supervision instead of labels in on-device training One of the primary challenges to real-world federated learning is the lack of supervised labels. Due to the no-peek nature of on-device data, manual labeling of individual training examples by raters is impossible. Given this constraint, three common approaches to federated training attempt to leverage the abundance of unlabeled data: unsupervised learning, semi-supervised learning, and supervised learning with contextual on-device signals. Several early applications of federated learning in production environments relied on unsupervised training (for example, the RNN language model in [104]). Techniques such as distillation or noisy student training can be used to build a student model with federated learning that outperforms an original teacher model [105]. And rater-assigned labels that are used for traditional server-based training can be replaced in certain circumstances with contextual signals. For example, accepted suggestions or manual corrections by users can be interpreted as supervision signals.

Use federated analytics to characterize client data and model behavior Federated analytics [92] can be used to get insights from federated data in a privacy-preserving way. These insights can improve the modeling workflow in federated settings. For example, federated analytics can be used to measure the fraction of users who type a particular token for a language modeling task for mobile keyboard applications, which can be useful for deciding whether the token should be part of the model’s input vocabulary. Federated aggregation can also be applied on the model’s outputs to better understand model behavior (a simple extension of federated aggregation of metrics like loss and accuracy).

Use generative models to characterize client data Augenstein et al. [17] propose using differentially private federated generative models to generate realistic samples of client data. This can in principle be useful for debugging and generating and testing model hypotheses.

6 Federated Optimization Theory

In this section, we will briefly discuss possible theoretical analysis for the convergence of current federated learning algorithms. In particular, we first introduce some theoretical tools for the convergence analysis in Section 6.1 by presenting a simple proof for the vanilla FEDAVG. Based on the simple analysis, we discuss the effects of local updates and data heterogeneity and compare it with well-known baselines in optimization theory. Later in Section 6.2, we review more recent literature to demonstrate how to relax the assumptions made for simplicity and improve over the basic results in Section 6.1.

6.1 Basic Convergence Analysis for Federated Optimization Algorithms

We present a simple proof for the vanilla FEDAVG [181], also known as Local SGD or parallel SGD for homogeneous data in the literature, to showcase the structure, tools, and tricks used for getting theoretical results in federated optimization. We begin by considering a simple setting of the generalized FEDAVG framework (Algorithm 1) and later on give pointers to literature for how things change in different settings. The techniques and insights provided in this subsection can be found in much previous literature, including [138, 142, 237, 260, 270, 271, 289].

6.1.1 Assumptions and Preliminaries

Formally, we make the following assumptions for the analysis. Specifically, assumptions (i-iv) are about the algorithmic choices in FEDAVG; assumption (v) and (vi) are about the properties of local objectives; assumption (vii) is about the data heterogeneity/similarity across clients.

- (i) At any round t , each client takes a total of τ local SGD steps with constant learning rate η , namely $\mathbf{x}_i^{(t,k+1)} = \mathbf{x}_i^{(t,k)} - \eta g_i(\mathbf{x}_i^{(t,k)})$, where $k \in [0, \tau)$ and g_i is the stochastic gradient at client i .
- (ii) The SERVEROPT takes a unit descent step with server learning rate 1.0, namely $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \Delta^{(t)}$.
- (iii) There are finite (M) clients and each client contributes a uniform share of the global objective. We label the clients as $\{1, 2, \dots, M\}$. Formally, we have $F(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M F_i(\mathbf{x})$.
- (iv) Each client participates *every round*, namely $\mathcal{S}^{(t)} \equiv \{1, 2, \dots, M\}$.
- (v) Each local objective $F_i(\mathbf{x})$ is *convex* and *L-smooth*.
- (vi) Each client can query an *unbiased* stochastic gradient with σ^2 -*uniformly bounded variance* in ℓ_2 norm, namely

$$\mathbb{E}[g_i(\mathbf{x}_i^{(t,k)}) | \mathbf{x}_i^{(t,k)}] = \nabla F_i(\mathbf{x}_i^{(t,k)}), \quad \mathbb{E}[\|g_i(\mathbf{x}_i^{(t,k)}) - \nabla F_i(\mathbf{x}_i^{(t,k)})\|^2 | \mathbf{x}_i^{(t,k)}] \leq \sigma^2. \quad (11)$$

- (vii) The difference of local gradient $\nabla F_i(\mathbf{x})$ and the global gradient $\nabla F(\mathbf{x})$ is ζ -uniformly bounded in ℓ_2 norm, namely

$$\max_i \sup_{\mathbf{x}} \left\| \nabla F_i(\mathbf{x}_i^{(t,k)}) - \nabla F(\mathbf{x}_i^{(t,k)}) \right\| \leq \zeta. \quad (12)$$

In a typical convergence proof we want to show that our iterates $\mathbf{x}^{(t,k)}$ is taking the objective function value $F(\mathbf{x}^{(t,k)})$ closer to the optimal value $F(\mathbf{x}^*)$. In the federated learning setup, we have multiple local iterates on clients and we want to ensure all of them are approaching the minimizer. To handle iterates from multiple clients, a concept of shadow sequence is introduced (commonly used in and originating from decentralized optimization literature [167, 293]), which is defined as $\bar{\mathbf{x}}^{(t,k)} := \frac{1}{M} \sum_{i=1}^M \mathbf{x}_i^{(t,k)}$. Given this notation, we have

$$\bar{\mathbf{x}}^{(t,k+1)} = \bar{\mathbf{x}}^{(t,k)} - \frac{\eta}{M} \sum_{i=1}^M g_i(\mathbf{x}_i^{(t,k)}). \quad (13)$$

From (13), observe that the averaged iterates actually perform a perturbed stochastic gradient descent. Its gradients are evaluated at client models $\mathbf{x}_i^{(t,k)}$ instead of $\bar{\mathbf{x}}^{(t,k)}$. We want to show that:

$$\mathbb{E} \left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{\mathbf{x}}^{(t,k)}) - F(\mathbf{x}^*) \right] \leq \text{an upper bound decreasing with } T. \quad (14)$$

How fast the above quantity decreases with T is called the *rate of convergence*. It is also worth noting that at the end of each round, we have $\bar{\mathbf{x}}^{(t,\tau)} = \bar{\mathbf{x}}^{(t+1,0)} = \mathbf{x}^{(t+1)}$. So the above quantity (14) also quantifies the convergence of the global model.

6.1.2 Main Results

We introduce two lemmas to facilitate the proof.

1. Show that we are making progress in each round, i.e. $\mathbb{E} \left[\frac{1}{\tau} \sum_{k=1}^{\tau} F(\bar{\mathbf{x}}^{(t,k)}) - F(\mathbf{x}^*) \right]$ is bounded by the difference of a potential function evaluated at t and $t+1$, measuring the progress of optimization, plus an additional small error term.
2. Show that all client iterates remain close to the global average/shadow sequence, i.e. in expectation $\|\mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)}\|^2$ remains bounded.

The first lemma is similar to analysis strategy for centralized optimization algorithms, but now we have added responsibility to ensure no client is deviating away. Once we have these two, by simple telescoping over t we can show that over T rounds we have made significant progress towards the optimal value. We formally state the two lemmas below, and leave elaborate proof of the lemmas to Appendix D.

Lemma 1 (Per Round Progress). *Assuming the client learning rate satisfies $\eta \leq \frac{1}{4L}$, then*

$$\mathbb{E} \left[\frac{1}{\tau} \sum_{k=1}^{\tau} F(\bar{\mathbf{x}}^{(t,k)}) - F(\mathbf{x}^*) \middle| \mathcal{F}^{(t,0)} \right] \leq \underbrace{\frac{1}{2\eta\tau} \left(\|\bar{\mathbf{x}}^{(t,0)} - \mathbf{x}^*\|^2 - \mathbb{E} \left[\|\bar{\mathbf{x}}^{(t,\tau)} - \mathbf{x}^*\|^2 \middle| \mathcal{F}^{(t,0)} \right] \right)}_{\text{Progress}} + \underbrace{\frac{\eta\sigma^2}{M} + \frac{L}{M\tau} \sum_{i=1}^M \sum_{k=0}^{\tau-1} \mathbb{E} \left[\|\mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)}\|^2 \middle| \mathcal{F}^{(t,0)} \right]}_{\text{Deviation}},$$

where $\mathcal{F}^{(t,0)}$ is the σ -field representing all the historical information up to the start of the t -th round.

Lemma 2 (Bounded Client Drift). *Assuming the client learning rate satisfies $\eta \leq \frac{1}{4L}$,*

$$\mathbb{E} \left[\|\mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)}\|^2 \middle| \mathcal{F}^{(t,0)} \right] \leq 18\tau^2\eta^2\zeta^2 + 4\tau\eta^2\sigma^2,$$

where $\mathcal{F}^{(t,0)}$ is the σ -field representing all the historical information up to the start of the t -th round.

Combine Lemmas 1 and 2 and telescope t from 0 to $T-1$ to obtain the main theorem as follows.

Theorem 1 (Convergence Rate for Convex Local Functions). *Under the aforementioned assumptions (a)-(g), if the client learning rate satisfies $\eta \leq \frac{1}{4L}$, then one has*

$$\mathbb{E} \left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{\mathbf{x}}^{(t,k)}) - F(\mathbf{x}^*) \right] \leq \frac{D^2}{2\eta\tau T} + \frac{\eta\sigma^2}{M} + 4\tau\eta^2L\sigma^2 + 18\tau^2\eta^2L\zeta^2, \quad (15)$$

where $D := \|\mathbf{x}^{(0,0)} - \mathbf{x}^*\|$. Furthermore, when the client learning rate is chosen as

$$\eta = \min \left\{ \frac{1}{4L}, \frac{M^{\frac{1}{2}}D}{\tau^{\frac{1}{2}}T^{\frac{1}{2}}\sigma}, \frac{D^{\frac{2}{3}}}{\tau^{\frac{2}{3}}T^{\frac{1}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}, \frac{D^{\frac{2}{3}}}{\tau T^{\frac{1}{3}}L^{\frac{1}{3}}\zeta^{\frac{2}{3}}} \right\}, \quad (16)$$

we have

$$\mathbb{E} \left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{\mathbf{x}}^{(t,k)}) - F(\mathbf{x}^*) \right] \leq \underbrace{\frac{2LD^2}{\tau T} + \frac{2\sigma D}{\sqrt{M\tau T}}}_{\text{Synchronous SGD}} + \underbrace{\frac{5L^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{\tau^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{19L^{\frac{1}{3}}\zeta^{\frac{2}{3}}D^{\frac{4}{3}}}{T^{\frac{2}{3}}}}_{\text{Add'l errors from local updates \& non-IID data}}. \quad (17)$$

6.1.3 Discussion

Theorem 1 can be used to obtain insights for FEDAVG or local-update algorithms. Below, we discuss the implications of Theorem 1 in detail. Similar conclusions can be found in previous literature [138, 142, 237, 260, 270, 271, 289].

Effects of local steps An important benchmark to compare with is *synchronous SGD* that synchronizes local models at each local iteration (that is, setting the number of local steps τ to be 1 in FEDAVG). In synchronous SGD, the last two terms in (15) and (17) will not appear at all and the rate of convergence is $\mathcal{O}(1/\tau T + \sigma/\sqrt{M\tau T})$. However, when clients perform multiple local steps, there will be an additional error in the upper bound (*i.e.*, last two terms in (15) and (17)). Fortunately, observe that the additional error terms is in proportion to η^2 and can decay in a rate of $\mathcal{O}(1/T^{\frac{2}{3}})$ (when the learning rate is small enough according to (16)). When the total communication rounds T is sufficiently large and the stochastic noise is not zero ($\sigma^2 \neq 0$), performing local steps will not degrade the rate of convergence.

Savings in communication rounds Taking local steps can save total communication rounds compared to synchronous SGD. To be more specific, when the total number of gradient evaluations/computations across all clients ($K = M\tau T$) is fixed and the local steps τ satisfies

$$\tau \leq \min \left\{ \frac{\sigma}{DL} \frac{K^{\frac{1}{2}}}{M^2}, \frac{\sigma}{\zeta} \sqrt{\frac{\sigma}{DL} \frac{K^{\frac{1}{2}}}{M^2}} \right\}, \quad (18)$$

the error upper bound (17) will be dominated by the second term $\mathcal{O}(1/\sqrt{K})$, which is the same as synchronous SGD. On the other hand, the local-update algorithm only takes T communication rounds within K parallel SGD iterations, while synchronous SGD costs τT communication rounds. It is obvious that when the upper bound of local steps (18) becomes larger, there will be more communication savings. Therefore, the quantity (18) represents the largest savings in communication rounds and is reported by many previous works.

Effects of data heterogeneity Theorem 1 also depicts that the data heterogeneity exacerbates the side effects of local updates on the convergence. Note that the additional error terms in (15) is in an order of $\mathcal{O}(\tau^2)$. However, if all local objectives are identical so that $\zeta = 0$, then the additional error terms will be just linear to the number of local steps. Besides, from (18) we know that, in the presence of high data heterogeneity ($\zeta \gtrsim \sigma$), the largest number of local steps is $\tau = \mathcal{O}(K^{\frac{1}{4}} M^{-1})$, which is much smaller than the IID data distribution case ($\tau = \mathcal{O}(K^{\frac{1}{2}} M^{-2})$).

Remark 2. While the above discussions around Theorem 1 are mainly under the stochastic setting ($\sigma \neq 0$), the conclusions may change in the deterministic setting (*i.e.*, clients use full-batch gradients to perform local updates and $\sigma = 0$). In particular, the rate of convergence under heterogeneous data setting will be substantially slowed down to $\mathcal{O}(1/T^{\frac{2}{3}})$ from $\mathcal{O}(1/\tau T)$.

Comparison with large-batch synchronous SGD So far, all the above discussions are based on the comparison with synchronous SGD. It is worth noting that there is another natural baseline to compare with. In particular, if we force all clients to use τ -fold larger mini-batch size in synchronous SGD algorithm, then this algorithm can also save τ times communication, which is the same as local-update algorithms. We refer to this algorithm as *large-batch synchronous SGD*⁵. It has been

⁵Both large-batch and regular synchronous SGD are equivalent to mini-batch SGD. In order to distinguish the different batch sizes ($M\tau$ versus M) and avoid confusions, we refer to them with different names.

well understood (see [66]) that the worst-case error of large-batch synchronous SGD is:

$$\Theta\left(\frac{LD^2}{T} + \frac{\sigma D}{\sqrt{M\tau T}}\right). \quad (19)$$

Now we are going to compare (19) and (17) in the regime where the statistical term $\mathcal{O}(1/\sqrt{M\tau T})$ does not dominate the rate⁶. The key insight is that FEDAVG or local-update algorithms are certainly not universally better, but there do exist some regimes where they can improve over large-batch synchronous SGD. For example, when the data heterogeneity is very low such that $\zeta^2 \lesssim 1/T$, the worst case error bound of FEDAVG (17) will be smaller than that of large-batch synchronous SGD (19). Similar conclusions hold for data homogeneous case (*i.e.*, when $\zeta = 0$), in which FEDAVG’s guarantee (17) can be better than (19) when the number of local steps is sufficiently large $\tau \gtrsim T$. Moreover, in the homogeneous setting, it has been shown in [271] that under possibly restricted conditions, such as Hessian smooth loss functions, FEDAVG can be strictly better than large-batch synchronous SGD.

Lower bounds Looking at lower bounds can enlighten us about what aspects can be improved and what cannot. For example, [138] derived lower bounds showing that heterogeneity can be particularly problematic, showing degradation as heterogeneity increases. Putting this in light of above comparisons to large-batch synchronous SGD, the lower bounds of [138] show that (17) cannot be substantially improved, and thus FEDAVG can be strictly inferior to large-batch synchronous SGD when the level of data heterogeneity is large or the number of local steps is small (in the data homogeneous setting) [270]. Further insights into problems caused by heterogeneity can be obtained by looking at the communication lower bound of [14], which matches the upper bound of accelerated gradient descent in the general heterogeneous setting. So, in order to achieve any nontrivial improvements in communication complexity, one has to introduce additional conditions on the second order heterogeneity, as considered in the distributed computing literature (without device sampling) by [210, 226, 300] and in federated learning by [137, 138].

6.2 Advances in Convergence Guarantees

The convergence property of FEDAVG (or local-update algorithms) has been extensively investigated in recent years. One can relax some assumptions in Section 6.1 via more complex analysis. In this subsection, we briefly discuss some recent advances in convergence guarantees and how to possibly improve over the results in Section 6.1.

6.2.1 Assumptions on data heterogeneity

The similarity, or dissimilarity, between the functions in (1) is an important quantity that determines the difficulty of the problem, both in practice and in theory. The challenge for theoretical analysis can be seen through a simple thought experiment. Consider a hypothetical situation where all clients own exactly the same set of training data. In this situation, from the perspective of any client, nothing could be gained from others’ data, and hence there is no need for communication among the clients. On the other hand, in the hypothetical situation where all devices have IID samples from the same underlying distribution \mathcal{D} , there is incentive to learn jointly in a federated fashion as there is more statistical power in the union of these IID datasets, especially when the number of devices is large. However, this IID regime is still not particularly challenging as the high level of data homogeneity inherent in this regime can be exploited. The most challenging regime, and the one which often occurs in practice, is when the clients contain data samples from widely different distributions.

⁶Otherwise, both algorithms have the same worst-case performance.

In order to quantify the dissimilarity between client data distributions, various measures have been proposed. Most common are quantities that are based on the gradients, in particular, the global bound on the variance of the gradients, taking the form

$$\mathbb{E}_{i \sim \mathcal{P}} \|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \zeta^2 + \beta^2 \|\nabla F(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, \quad (20)$$

where ζ^2, β^2 are parameters. When $\beta = 0$, the above bound can recover assumption (vii) used in Section 6.1. But it has been shown that setting $\beta > 0$ can lead to refined rates [96, 137, 138, 147, 161, 261]. When $\zeta = 0, \beta \neq 0$, (20) recovers the assumption used in [161].

No data similarity Given that the client’s data are highly non-iid, it might happen that there is no finite β, ζ such that (20) holds. Such an issue has been recently addressed by Khaled et al. [142] given that problem (1) has an unique minimizer \mathbf{x}^* (as for instance for strongly convex functions, or non-convex functions under the Polyak-Lojasiewicz condition). Specifically, it is enough to impose the condition (20) only at a single point \mathbf{x}^* instead of the full domain \mathbb{R}^d . Similar to the observation of Theorem 1, a price is paid for worse communication guarantees even with this weaker assumption [137, 142, 147].

Other heterogeneity assumptions Li et al. [166] quantify the data heterogeneity based on the quantity $F^* - \mathbb{E}_{i \sim \mathcal{P}}[F_i^*]$, where F^* and F_i^* are the minimum values of F and F_i respectively. Zhang and Li [295] assume—motivated by over-parametrized problems—that the intersection of the solution sets on the clients is nonempty. Other works assume conditions on the Hessians [14, 137, 138, 226] or assume that the deviation between each client’s and the global objective is smooth [42]. Given the strong lower bounds known for the performance of local update methods under the standard setting of (20), exploring such alternate heterogeneity assumptions is an important direction to bridge the gap between the theoretical and practical performance of the methods.

6.2.2 Extensions and Improvements

Analyses for various objective functions Instead of assuming the local objective function to be general convex, one can extend the results for strongly convex local objectives, see [96, 138, 142, 147, 237, 270]. In the strongly convex setting, the additional errors in (17) will decay faster along with the communication rounds T . We defer interested readers to Table 2 in [270] that summarizes the current state-of-the-art results. There is much literature analyzing non-convex settings, the convergence rates (in terms of $\mathbb{E}[\|\nabla F(\mathbf{x}^{(t)})\|^2]$) of which are similar to the general convex setting in Theorem 1, see [138, 147, 261].

Analyses with client sampling Analyzing FEDAVG with client sampling under certain constraints can generally be done by standard techniques. In particular, when using client sampling, the update rule of the virtual sequence (13) can be rewritten as follows

$$\bar{\mathbf{x}}^{(t,k+1)} = \bar{\mathbf{x}}^{(t,k)} - \eta \sum_{i \in \mathcal{S}^{(t)}} w_i^{(t)} g_i(\mathbf{x}_i^{(t,k)}) \quad (21)$$

where $w_i^{(t)}$ can be chosen to ensure that $\mathbb{E}_{\mathcal{S}}[\sum_{i \in \mathcal{S}} w_i^{(t)} g_i(\mathbf{x}_i^{(t,k)})] = \sum_{i=1}^M g_i(\mathbf{x}_i^{(t,k)})/M$ so that the client sampling scheme just adds a zero-mean noise to the update rule (13) and the convergence of FEDAVG can be naturally guaranteed. In the resulting error bound, the number of total clients M will be replaced by the number of active clients per round $|\mathcal{S}^{(t)}|$. Interested readers can refer to [96, 138, 161, 166, 211, 261, 283] for the technical details.

Analyses with advanced SERVEROPT and CLIENTOPT While the basic analysis in Section 6.1 assumes that the server takes a unit descent step to update the global model, one can replace it with other SERVEROPT. It has been shown in [137, 138, 211] that, with properly chosen server learning rate η_s for $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \eta_s \Delta^{(t)}$, the convergence rate can be improved. When the SERVEROPT is momentum SGD or adaptive optimization methods, Reddi et al. [211] proves that they preserve the same convergence rate as vanilla FEDAVG but with significantly improved empirical performance. In another line of research, using momentum or accelerated SGD or adaptive optimizers on clients has been explored in [137, 261, 263, 288, 291]. In particular, Karimireddy et al. [137], Yuan and Ma [291] proposed provably accelerated variants of FEDAVG.

Analyses with communication compression As discussed in Section 3.1.2, combining FEDAVG with compression can potentially boost the communication efficiency. The theoretical analysis without local steps was explored in [10, 239], and with local steps in [30]. More recently there is much more literature studying communication compression, the update rule of which can be written as $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \sum_{i=1}^M p_i \mathcal{C}_i^{(t)}(g_i(\mathbf{x}^{(t)}))$ where $\mathcal{C}_i^{(t)}$ denotes the suitably chosen compressor. In order to guarantee the convergence, certain constraints have to be enforced on the class of compressors. Alistarh et al. [8], Horváth et al. [117, 118] studied unbiased compressors that satisfy $\mathbb{E}[\mathcal{C}(\mathbf{x})] = \mathbf{x}$ and $\mathbb{E}[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2] \leq \omega \|\mathbf{x}\|^2$ for some $\omega \geq 0$ and all $\mathbf{x} \in \mathbb{R}^d$. The theoretical analysis of unbiased compressor is by adding a zero-mean noise to the stochastic gradients. Moreover, biased compressors can be analyzed, for instance contractive compressors that satisfy $\mathbb{E}[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2] \leq (1 - 1/\delta) \|\mathbf{x}\|^2$ for some $\delta \geq 1$ and all $\mathbf{x} \in \mathbb{R}^d$. If applied naively, contractive compressors can diverge, even on simple low-dimensional piece-wise linear or quadratic problems [35, 136]. A practical fix was suggested by Seide et al. [225] without theoretical guarantees, which is known under various names including *error compensation*, *error feedback* and *memory feedback*. Error compensation remained without any theoretical support until recently [10, 136, 239, 272], with a steady stream of works achieving gradual progress in the theoretical understanding of this mechanism [35, 97, 146, 205, 206, 238, 247].

Analyses with control variates As discussed in Section 3.2, control variates methods have been proposed to explicitly reduce the client shift problem [137, 138]. The basic idea is to debias the local model updates on each client using a global estimation over multiple clients, similar to what was used in communication efficient distributed computing [226]. Such a bias correction addresses the client shift problem, leading to better convergence guarantees for local optimization methods.

Specifically, *without any assumptions on the data heterogeneity*, the SCAFFOLD algorithm developed in [138] converges at a rate of $O(1/M\tau T) + O(e^{-T})$ for strongly convex functions and $O(1/\sqrt{M\tau T})$ for non-convex functions. SCAFFOLD, under second order regularity assumptions, shows strictly better rates than mini-batch SGD matching the lower-bound of [14]. Similarly, when the MIME framework developed in [137] is instantiated with momentum-based variance reduction, the resulting algorithm obtains the optimal convergence rate of $O((M\tau T)^{-2/3})$ for non-convex functions assuming that the clients' gradients and Hessians satisfy certain closeness properties. This algorithm provably outperforms both large-batch synchronous SGD and FEDAVG in the cross-device setting, matching the lower bound of [15].

6.2.3 Gaps Between Theory and Practice

In this subsection we discuss some of the pressing gaps between theoretical analyses of federated optimization algorithms, and practical algorithmic performance. We note that this is a critical topic, even in non-federated machine learning. There are a number of works dedicated to bridging the gap between the theory and practice of machine learning, especially deep learning, often focusing on topics such as generalization and the empirical success of optimization methods in non-convex

settings (see [31, 46, 130, 235] for a few examples of such work, which we do not cover in this work).

We focus on gaps between theory and practice that are endemic to, or at least exacerbated by, federated learning. We stress that federated learning can be difficult to analyze theoretically due to its heterogeneous nature. Therefore, the topics discussed below are *aspirational*, and should be viewed as significant open problems in the field. Moreover, this is not a complete list of open problems; some of the topics in Section 3 for practical consideration can become restrictions for theory. For example, algorithms that do not maintain persistent state and have low cost computation on clients are preferred for cross-device FL. For more details on such open problems in this vein, see [132, Section 3].

Asymptotic versus communication-limited performance One important caveat to many of the convergence rates discussed above is that they are *asymptotic*, and may require the number of communication rounds T to be sufficiently large in order to become valid. Such asymptotic results can belie the practical importance of an algorithm in settings where T is small. This communication-limited nature was part of the motivation given by McMahan et al. [181] for the development of FEDAVG. As McMahan et al. [181] show, FEDAVG can lead to drastic improvements over algorithms such as FEDSGD (another name for large-batch synchronous SGD discussed in Section 6.1) in regimes where T is limited, despite the fact that FEDAVG and FEDSGD often have comparable asymptotic convergence rates [270]. Empirical examples of how communication-limited regimes can alter the relative performance of federated algorithms were given in Section 4. Recent work suggests that empirically successful federated optimization algorithms such as FEDAVG and FEDPROX can quickly arrive in neighborhoods of critical points, but may not actually converge to such critical points [43, 177]. It remains an open question how to theoretically quantify the communication-limited benefits of algorithms such as FEDAVG, especially in non-convex settings.

Biased sampling of clients Another common facet of many theoretical analyses of federated learning is the assumption that we can sample clients in an unbiased or close to unbiased fashion. As discussed by Bonawitz et al. [39] and Eichner et al. [75], practical cross-device federated learning usually must contend with non-uniform client availability. Whether clients are available to participate in a round depends on a number of factors (such as geographic location of the device, in the case of mobile phones). This presents significant challenges empirically and theoretically. Two outstanding challenges are understanding what types of biased sampling still result in good performance of existing algorithms [e.g. 57, 60], and how to develop algorithms that can better contend with biased client sampling. While this is related to work on shuffling data in centralized learning (see [220] for example) we stress that the degree of non-uniformity in the client sampling may be much greater in federated learning, especially if clients locally determine their availability for training [39].

Client aggregation methods In addition, another area where theory and practice are sometimes not aligned is in the client aggregation methods used. Recall that in Algorithm 1, the client updates are aggregated according to a weighted average, where client i is weighted by $p_i / \sum_{i \in \mathcal{S}^{(t)}} p_i$. The weights are *time-varying* depending on $\mathcal{S}^{(t)}$. Theoretical analyses [e.g. 166] often use $p_i = 1$ for all i in order to reduce this aggregation to simple averaging across clients. However, as discussed by McMahan et al. [181], setting the weights to be in proportion to the number of examples in the local dataset of client i may be more fruitful, as it allows one to recover a global loss function that is an average across all examples. The gap between this uniform aggregation and example-weighted time-varying aggregation can be especially important for extremely unbalanced datasets, where some clients may have only a few examples, while others may have thousands. Thus, deriving convergence theory that analyzes the performance of such aggregation methods in realistic settings could be a valuable area of exploration.

7 Privacy, Robustness, Fairness, and Personalization

In this section, we discuss the relationship between federated optimization and other important considerations, such as privacy, robustness, fairness, and personalization. These problems are often posed as explicit or implicit constraints in the objective function used during optimization, as well as within the dynamics of federated optimization methods. We do not intend to give a comprehensive review of these topics. Our intention is to highlight how these added constraints impact federated optimization. For a more in-depth discussion of open problems, see sections 4-6 in [132].

7.1 Privacy

While federated learning does not permit the server to access clients' data directly, federated optimization methods compute aggregate model updates using clients' data, and send these to the server. While these updates should not directly contain clients' data, it is inevitable that they contain downstream information about the clients' datasets. In fact, Zhu and Han [301] show that it is possible in some cases to reconstruct client data to some extent from model gradients. Thus, it is critical that federated optimization methods are developed with privacy in mind. In order to aid the development of privacy-first methods, we describe two general techniques for privacy that are often used in federated optimization.

7.1.1 Data Minimization

One class of technologies often used to strengthen privacy by reducing the surface of attack is generally referred to as data minimization. In federated learning, a default level of data minimization is achieved by ensuring that raw data is never shared with the server and that only focused, minimal updates that are intended for immediate aggregation are sent back to the server.

In order to achieve greater privacy levels, we often use additional data minimization techniques. One such technique is secure computation. Secure computation provides cryptographic primitives that enable federated aggregation without revealing raw inputs to untrusted parties. [32, 38] propose secure aggregation (SECAGG), which enables computing secure federated sums over client inputs. This permits the calculation of (weighted) averages of client updates without revealing individual updates to the server.

While SECAGG enforces a greater level of privacy, it also imposes constraints and limitations that must be accounted for in the development of SECAGG-compatible optimization methods. One notable constraint is that SECAGG operates on the level of federated sums (*i.e.*, sums of client updates). Algorithms that require aggregations that cannot be performed through federated sums are therefore incompatible. While potentially restrictive, this constraint opens up the door to novel research on how to develop SECAGG-compatible versions of federated optimization methods. For example, Pillutla et al. [201] shows that median-based aggregation can be well-approximated by a small number of federated sums (on the same set of clients), and Karimireddy et al. [139], Vogels et al. [253] design new communication compression and robust aggregation protocols specifically compatible with SECAGG. It remains an open question how to approximate other aggregation strategies (such as model distillation and fusion [173]) using only federated sums, and how to implement practical secure computation method beyond summation for complicated aggregation methods.

7.1.2 Data Anonymization

Another class of technologies used to strengthen privacy is data anonymization. Such methods seek to introduce plausible deniability into the output of an algorithm. One such method is *differential*

privacy. Differential privacy is notable for its rigorous nature and well-tested formalization of the release of information derived from sensitive data. Informally, a differentially private mechanism is a randomized algorithm for a database query with a quantifiable guarantee that it will be difficult for an adversary to discern whether a particular item is present in the data, given the output of the query and arbitrary side information. The privacy guarantee is quantified by two values, ϵ and δ , with smaller values implying increased privacy. In a machine learning context, the “query” is an algorithm (such as SGD) that operates on training data and outputs a model. In centralized settings, many differentially private learning algorithms define the unit of privacy to be a single training example. However, in federated settings it is often preferable to guarantee privacy with respect to each users’ entire data collection [180]—an adversary should not be able to discern whether a user’s dataset was used for training, or anything about it.

In the context of cross-silo FL, the unit of privacy can take on a different meaning. For example, it is possible to define a privacy unit as all the examples on a data silo if the participating institutions want to ensure that an adversary who has access to the model iterates or final model cannot determine whether or not a particular institution’s dataset was used in the training of that model. User-level DP can still be meaningful in cross-silo settings where each silo holds data for multiple users. However, enforcing user-level privacy may be challenging if multiple institutions have data from the same user. The conventional per data point privacy may be of interests in a lot of cross-silo settings.

In the context of cross-device FL, federated optimization provides a natural way to give this type of “user-level” differential privacy guarantee. A client’s contribution to a training round is summarized by its update: the data that is sent from clients to server at each round. By applying privatizing transformations to these updates, we can derive user-level DP guarantees. For example, in the case of FEDAVG, the client updates can be clipped (bounding their ℓ_2 norm) and a calibrated amount of Gaussian noise can be added to the average, which is often sufficient to obscure the influence of any single client [180].

Central versus local privacy Sharing the global model between clients or any external adversary could reveal sensitive information about the clients’ data [1, 48]. We consider two settings of privacy. The first is called the *central privacy* setting, where a trusted server can collect model updates from the clients, clip and aggregate the model updates, and then add noise to the aggregate before using it to update the server-side model and sharing the updated model with (possibly) untrusted parties (or clients in later federated rounds). Hence, this setting of differential privacy protects the privacy of the clients’ data against external parties by relying on a trusted server. The second setting is called *local privacy*, where each client preserves the privacy of their own data against any (possibly) compromised entity including the central server, which could be untrusted. Here, the privacy guarantee is with respect to the server as well as other clients participating in the distributed optimization. This is achieved by clipping and noising the model updates locally (on device) before sending the noised and clipped model updates back to the server.

The local privacy setting provides a stronger notion of privacy against untrusted servers; however, it suffers from poor learning performance in comparison with the central privacy setting because significantly more noise must be added to obscure individual updates [72, 131, 140]. Fortunately, local privacy can be boosted when combined with the use of either a *secure shuffler* (that receives the private local updates of the participated clients and randomly permutes them prior to sending them to the untrusted server [22, 76, 85, 90]) or a *secure aggregator* that sums the private updates of the participated clients prior to communicating the aggregate update to the untrusted server [134]. Observe that secure shuffling/aggregation plays the role of anonymization in which the server that observes the aggregate update cannot assign a certain update to a specific client, which can amplify the privacy of the local privacy model. Hence, the secure shuffler/aggregator can help us design

private learning algorithms that have a reasonable learning performance without assuming the existence of a fully trusted server. These approaches are often considered distributed DP, and the mathematical privacy guarantee are expressed as central privacy bound.

The impact of private training on optimization Many recent works have studied the impact of differentially private mechanisms on centralized optimization [26, 27, 28, 61, 81, 90, 231]. Such works are often concerned with developing better differential privacy mechanisms for model training, and for getting tight bounds on the (ϵ, δ) privacy guarantees of such mechanisms. One particularly important method is DP-SGD [26], which applies SGD, but adding properly calibrated noise to the gradients. Such work has since been extended to differentially private distributed optimization algorithms [76, 90, 140], including differentially private federated learning [180]. Much of the work in this area reveals fundamental *privacy-utility trade-offs*. Generally speaking, to guarantee stronger privacy with smaller (ϵ, δ) , the utility (*e.g.*, accuracy in a recognition task) is often sacrificed to a certain level. While typically unavoidable (in formal, information-theoretic ways), one can often mitigate this trade-off and provide stronger privacy by introducing properly designed differential privacy and randomization mechanisms.

We note that developing and analyzing differentially private federated optimization methods is a relatively new area of study, and there are a number of open questions around how to combine the optimization techniques, such as those discussed in Section 3.2, with differentially private mechanisms. We list a few of the challenges below.

Unknown, unaddressable, and dynamic population Providing formal (ϵ, δ) guarantees in the context of cross-device FL systems can be particularly challenging because the set of all eligible users is dynamic and not known in advance, and the participating users may drop out at any point in the protocol. There are few works discussed this issue: the recent work of Balle et al. [23] shows that these challenges can be overcome using a novel random check-ins protocol (*i.e.*, clients randomly decide whether to and when to check in without any coordination); [89] analyzes DP-SGD with alternate client self-sampling; the more recent work of [135] proposes a new online learning based DP algorithm, differentially private follow-the-regularized-leader (DP-FTRL), that has privacy-accuracy trade-offs that are competitive with DP-SGD. DP-FTRL does not require amplification via sampling or a fixed, known population. Building an end-to-end protocol that works in production FL systems is still an important open problem.

Variance in differential privacy methods Conventional differentially private optimization algorithms can increase the variance of the stochastic gradient method to the order of d , where d is the dimension of the model. Since modern machine learning models may have millions or even billions of parameters, this scaling with d can lead to significantly less accurate models. This issue can become a bottleneck for federated learning under privacy constraints, making it difficult to privately learn large network with reasonable accuracy. Some possible mitigation strategies include augmenting the learning procedure with public data that does not have any privacy concerns [29] or developing new private optimization methods with dimension-independent variance under certain assumptions regarding the objective function and stochastic gradients [128, 133]. In practical centralized learning setting, we can reduce the variance by increasing the number of samples at each round. In federated learning, this is analogous to increasing the number of clients participating in each round. While this may be possible in some setting (such as cross-device settings with high levels of client availability) it may not be possible in all settings (such as cross-silo settings with only a small number of clients).

Communication and privacy As discussed in Section 3, communication-efficiency is typically an important requirement for federated optimization algorithms. While various communication-

efficient optimization methods are discussed in Section 5.1, such methods may not preserve privacy. Furthermore, addressing privacy and communication-efficiency separately may lead to suboptimal methods. This has led to a budding area of communication-efficient private optimization algorithms. For example, Agarwal et al. [4], Kairouz et al. [134] propose a private optimization method where clients only need to send $\mathcal{O}(d)$ bits to the server at each round. Chen et al. [56], Girgis et al. [90] studied communication-efficient and privacy-preserving mean estimation with each of them giving optimal methods for various ℓ_p geometries. In particular, Girgis et al. [90] proposed and studied a variant of SGD that requires $\mathcal{O}(\log(d))$ bits per client, showing that the same privacy-utility as full precision gradient exchange can be achieved. Hence, addressing the problems of communication and privacy together can lead to learning algorithms with a significant reduction on the communication cost per client.

7.2 Robustness

Machine learning can be vulnerable to failures – from benign data issues such as distribution shift [155] to adversarial examples [16] and data poisoning [236]. Beyond these centralized robustness concerns, federated optimization, where hundreds of millions of client devices participate in the learning process with their local data, exposes the learning system to additional threats and novel versions of existing threats – since, unlike centralized settings, the central server has limited control of the system devices and communication channels. Therefore, robustness is essential for the trustworthy, practical deployment of federated learning.

In this section, we describe some of the possible adversarial attacks in federated learning, and survey existing techniques for defending against these attacks. We also discuss other important aspects of robustness, including robustness due to non-malicious occurrences and its relation to privacy. While federated learning introduces new vulnerabilities to inference-time attacks (see [132, Section 5.1.4] for a survey), we will focus on training-time attacks in our discussions below.

7.2.1 Goals of an Adversary

As discussed by [132, Section 5.1.1], there are a number of goals a training-time adversary may have. They may wish to reduce the global accuracy of a model (*untargeted attack*), or they may wish to alter the model’s behavior on a minority of examples (*targeted attacks*). For example, an adversary could attempt to make a model misclassify all green cars as birds while maintaining its accuracy on all other samples [21, 244, 256, 274]. Due to the average performance maintenance, the presence of a targeted attack can be difficult to detect. However, recent work has shown that untargeted attacks can also be introduced in ways that are difficult to detect from a statistical viewpoint [25, 277].

Adversaries may not always actively alter the training procedure. Instead, they may function as eavesdroppers, and attempt to glean sensitive information about clients using information available at the server or by simply using the final model. This can include reconstructing client data from model updates [267], or performing label-inference attacks on aggregated model updates [265]. While such instances are important failures of robustness, they may be better addressed via the privacy-preserving techniques discussed above. We will therefore focus on training-time attacks that attempt to alter the learned model in the following.

7.2.2 Model Poisoning and Data Poisoning Attacks

Since federated learning is typically a collaborative and iterative process, and the final model is often deployed on edge devices for inference, there are a number of ways in which an adversary may attack a federated learning system. One particularly important distinction is whether an adversary

is present during the training or during inference. In training-time attacks, the adversary corrupts a subset of clients and attacks the system either by corrupting the data at the compromised clients (*data poisoning attacks*), or by having the corrupt clients send spurious updates to the server during the learning process (*model poisoning attacks*). See [132, Section 5.1] for more a more thorough taxonomy and discussion of such attacks.

Rather than focusing on specific attacks, we discuss how data and model poisoning attacks differ, especially in federated contexts. While both can significantly mar the capabilities of a learned model, data poisoning attacks are special cases of model poisoning attacks, as the clients which have poisoned data report corrupted updates to the server. Thus, data poisoning attacks may be strictly weaker; while a single malicious client can compromise a entire learning process via model poisoning [37], significantly degrading model performance with data poisoning may require poisoning many clients [21]. While data poisoning can be strengthened with collaborative attacking [243] and defense-aware poisoning [36, 80], a full understanding of the difference in capabilities between data poisoning and model poisoning remains unknown. This is particularly important in federated learning. Data poisoning attacks are generally easier for an adversary to engage in, requiring only influencing the data collection process within a client, not actually circumventing the core operation of the federated learning process.

Another key notion is that data poisoning attacks are not necessarily the result of an explicit adversary. Data collection processes, especially in settings with heterogeneous clients, are often noisy and can inadvertently lead to outlier data. If data poisoning occurs before training begins, it may lead to inaccurate models, even in systems where clients train honestly, according to the specified algorithm. In contrast, model poisoning attacks are often modelled by assuming corrupt clients can send arbitrary and adversarially chosen vectors to the server throughout training. The dynamic nature of such attacks makes them significantly more challenging to safeguard against. It may be difficult to even detect such attacks in a distributed system, as the downstream failure of a model may have any number of possible explanations.

7.2.3 Defenses Against Training-Time Attacks

There are a number of proposed defenses against training time attacks, which can broadly be divided into two categories: coding-theoretic defenses and statistical defenses. Coding-theoretic solutions [52, 62, 65, 207] introduce redundancy across compute nodes in a distributed system, and do not require statistical assumptions on the data. Direct application of these methods in federated learning could require introducing redundancy across clients' local data, and may violate data minimization principles.

Statistical solutions typically require some kind of robust aggregation to mitigate the effect of corrupt updates. For example, one popular class of statistical defenses replaces the averaging step in SERVEROPT by a robust variant of the average such as the geometric median [58, 201], coordinate-wise median/trimmed-mean [275, 286], heuristics-based [37, 183], or other robust mean estimation techniques from statistics [63, 64, 242, 287]. Under assumptions on the distribution of the client updates and the number of corrupt clients, this can guarantee that the aggregate update is close to the true average.

Heterogeneous clients present a challenge to robust federated learning since it is hard to distinguish a client with unique data from a clever attacker [114]. Solutions for this setting attempt to re-use existing robust algorithms by either clustering the clients into relatively homogeneous clusters [86], or by properly resampling clients during training [114], and by using gradient dissimilarity properties [63, 64]. Such work often requires new theoretical approaches in order to contend with heterogeneous data, such as analyses incorporating non-uniform sampling [114] and novel concentration inequalities [63, 64]. A possible solution might be to use trusted data at the server to filter the client updates [276, 278].

However, trusted data is often not available at the server in federated learning and so such solutions are not feasible. Instead, the client updates can be directly utilized to build up estimates of non-outlier model updates over time, and memory-based solutions for combating adversarial attacks have been employed by [9, 11, 139]. These defenses however are currently only designed for the cross-silo setting and extending them to the cross-device setting is an important open question. Client heterogeneity also enables targeted attacks where an attacker may manipulate only a small subset of clients’ data in order to introduce backdoors [21]. Detecting and defending against such targeted attacks is challenging and remains an open problem [244, 256].

Another challenge in FL is that of communication efficiency, and only a few works have considered this with Byzantine adversaries [64, 88, 275]. Among these, Data and Diggavi [64] works with heterogeneous data and high-dimensional model learning, and incorporates local iterations in the analysis. However, the polynomial time method in [64] for filtering out corrupt updates is not as efficient as other cited methods. An important and interesting direction is to devise a communication efficient method that works with the challenges posed by the setting of federated learning.

7.2.4 Other Robustness Concerns

Robustness and aggregation protocols An existing challenge is to ensure that robust optimization is compatible with privacy primitives in federated learning such as SECAGG [38]. Many robust methods rely on non-linear operations such as medians, which are often not efficient when combined with cryptographic primitives. Some emerging solutions address this by replacing median operations with iterative Weizfeld’s algorithm approximation [201], or to clip the client updates before aggregation [139]. He et al. [115] uses secure two-party computation with a non-colluding auxiliary server in order to perform non-linear robust aggregation. In addition to restricting the non-linearity of defenses, aggregation protocols such as SECAGG also make training-time attacks more difficult to detect, as the server cannot trace suspected adversarial updates to specific clients [21]. Therefore, when designing a new defense for federated learning, we suggest considering its compatibility with secure aggregation protocols, and to be explicit about the type of system required for a method.

Robustness to distribution shift In heterogeneous settings where the client population evolves over time, it may be critical to train a model that is robust to distribution shift in data. In fact, in federated settings where we train a global model over some training set of clients and test its performance on a held-out set of clients, we are effectively measuring performance under some amount of distributional shift, as the two populations may be very different from one another. To this end, Lin et al. [173] uses ensembling throughout their federated training to produce more distributionally robust models. Another approach to distributional robustness involves having clients learn both a model and a “worst-case” affine shift of their data distribution [212].

Rate of corruption Many of the aforementioned works on defenses against training-time attacks provide resilience against that a constant fraction of clients are malicious (say, one-tenth of the clients are malicious). However, the effective rate of corruption may be extremely small in cross-device settings, where there may be hundreds of millions of clients, only a tiny fraction of which are actually malicious. In cross-silo settings, where there may be only a small number of clients, the rate of corruption may be higher. Given the already onerous difficulties of developing federated optimization methods that contend with heterogeneity, communication constraints, privacy, and compatibility with SECAGG, we emphasize that developing methods that can also tolerate a tiny fraction of malicious clients is still an important achievement, and may be a much more tractable problem.

7.3 Fairness

Algorithmic fairness is an emerging research area under machine learning that attempts to understand and mitigate the unintended or undesirable effects the learned models may have on individuals or sensitive groups (races, genders, religions, etc.) [24, 141]. For example, if individuals with similar preferences and characteristics receive substantially different outcomes, then we say that the model violates individual fairness [74]. If certain sensitive groups receive different patterns of outcomes, this can violate certain criteria of group fairness (e.g. demographic parity fairness or equality of opportunity fairness [106]).

Fairness concerns are critical, and can be exacerbated, in federated learning, due to systems and data heterogeneity. While models trained via federated learning may be accurate on average, these models can sometimes perform unfairly or even catastrophically on subsets of clients [162]. For instance, the unfairness/bias could come from periodic changes of data patterns, over-represented data due to the large size of samples on some specific devices, or under-represented users who do not own devices [132]. It is thus important to produce models that go beyond average accuracy to also ensure fair performance, e.g., maintaining a minimum quality of service for all devices.

This notion of fairness (i.e., uniformity of model performance distribution), also known as representation disparity [109], poses additional challenges for federated optimization. Some works propose min-max optimization [68, 184] to optimize the model performance under the worst case data distributions. As solving a min-max problem in federated settings can be especially challenging due to the scale of the network, other works propose alternative objectives to reweight samples less aggressively, allowing for a more flexible tradeoff between fairness and accuracy [162, 163]. Fairness can also be enforced by optimizing the dual formulation of min-max robust optimization, e.g., via superquantile methods or conditional Value-at-Risk which only minimizes the losses larger than a given threshold [153, 163]. Different from the motivation described above, there are also other works considering varying notions of fairness (e.g., proportional fairness [297]) for federated learning. We note that it is important for any fair federated learning algorithms to be resilient to heterogeneous hardware by allowing for low device participation and local updating, which are basic desirable properties for federated optimization methods, especially for cross-device settings where fairness concerns are prominent.

While not originally motivated by the problem of fairness, another line of work that can mitigate unfairness/sampling bias is alternate client selection strategies (as opposed to random selection) [57, 60, 69]. Intuitively, selecting representative devices could produce more informative model updates, thus helping speed up convergence and encouraging a more uniform performance distribution across all devices. However, discovering the underlying structure/similarities of the federated networks may be prohibitively expensive. Therefore, we suggest that any client selection methods need to account for and be compatible with practical systems and privacy constraints of FL [39].

There are still many open problems for fairness in FL. For instance, it would be interesting to explore the connections between the existing FL fairness notions and the broader fairness literature. It is also worth investigating the intersection between fairness and other constraints like robustness and privacy, as discussed in the next section.

7.4 Tensions Between Privacy, Robustness, and Fairness

It is also critical to consider the interplay and tension between fairness, robustness, and privacy. Fairness constraints attempt to ensure that the learned model works well for clients with data distributions that are not particularly well aligned with the average population distribution. Robustness constraints attempt to ensure that the output of a learning algorithm is not affected by outlier data. Privacy constraints attempt to ensure that the learned model and intermediate model updates do not

overfit to or retain information about outlier client training examples. Intuitively, these constraints can potentially lead to conflicting objectives.

For example, training with differential privacy requires bounding the clients’ model updates and adding noise to the model in each round, therefore preventing the model from memorizing unique training examples. Though this often helps with robustness [244], it leads to disparate impact on accuracy and therefore unfairness [20]. The use of secure aggregation to hide client updates strengthens privacy but often makes it impossible for the server to use robust optimization algorithms (e.g. computing a high dimensional median or a trimmed mean which require access to the individual model updates). Worse still, fairness and robustness are difficult to maintain while simultaneously ensuring privacy, as user privacy constraints limit the ability of a central provider to examine, test, and validate training data. On the other hand, robustness to distributional shift in user data distributions (discussed in Section 7.2) can be thought of as a generalized notion of fairness.

It is an interesting area of future research to optimize for fairness in federated learning while taking into account robustness and privacy constraints. As we will see in the next subsection, personalization might offer a good solution to break the tension between these constraints [164].

7.5 Personalization

As discussed in Section 1, a defining characteristic of federated learning is that client data is likely to be heterogeneous. One natural solution is to provide models that are *personalized* to each client. To enable personalization, a simple approach is to incorporate client-specific features. These features may be naturally occurring in the data, or may take the form of some auxiliary meta data. However, in lieu of (or in addition to) including such expressive features, we can use federated optimization methods to learn personalized models for clients. For example, Li and Wang [157] propose an algorithm that allows clients to have different model architectures via transfer learning, provided the server has access to a representative public dataset. However, for brevity, in this section we focus on the case when the personalized models have the same architecture.

The idea that every client gets a good model for its own data not only improves the overall statistical performance, but also potentially improve fairness or robustness of private algorithms [164, 290]. As alluded to in Section 7.4, using personalization to mitigate tension between privacy, robustness, and fairness is an active research area in federated learning. Before discussing the specific personalization algorithms (e.g., multi-task learning, clustering, fine-tuning, and meta-learning), we give two general categories of personalization algorithms:

- **Algorithms that require client-side state or identifier:** Here “client-side state” means that every client needs to maintain some state variables locally during training, and is responsible for carrying these variables from the previous round and updating them in the next round. As pointed out in Section 1.2, the assumption of stateful clients is more appropriate in the cross-silo setting, where the number of total clients is relatively small, and most (if not all) clients can participate in every round of training. Algorithms in this category sometimes also assume that every client has an identifier known by the server, and that the server can store updates from individual clients. This assumption can incur extra privacy concerns because it may be difficult to obtain strong privacy guarantees under this setting (see Section 7.1 for more discussions on privacy considerations in federated learning).
- **Algorithms that do not require client-side state or identifier:** In contrast to the first category, algorithms in this category do not require the server to know the client’s identifier; the clients also do not need to carry a state from the previous round to the next round. This makes these algorithms more attractive in the cross-device setting, where the population size is huge (e.g., millions of devices), only a small number of clients (e.g., a few hundreds) participate

in each round, and a device usually participates only once during the entire training process. Furthermore, it is often straightforward to combine these algorithms with the privacy techniques discussed in Section 7.1.

7.5.1 Algorithms that Require Client-side State or Identifier

A popular technique used in this category is **multi-task learning**⁷. To model the (possibly) varying data distributions on each client, it is natural to consider learning a separate model for each client’s local dataset. If we view learning from the local data on each client (or possibly a group of clients) as a separate “task”, we can naturally cast such a problem as an instance of multi-task learning. In multi-task learning, the goal is to learn models for multiple related tasks simultaneously. The ERM objective function usually takes the following form:

$$\min_{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in R^d} \sum_{i=1}^M f_i(\mathbf{x}_i) + \phi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M), \quad (22)$$

where \mathbf{x}_i and $f_i(\cdot)$ denote the local model and loss function for the i -th client, and $\phi(\cdot)$ is a regularization term capturing the relations between the clients.

Smith et al. [232] first proposed a primal-dual multi-task learning method for jointly learning the optimal local models and the relationships among these models in the federated setting. A similar formulation is considered in [101, 102, 124, 164], where the key differences are their regularization terms. The regularization term in [101, 102, 164] is the distance between the local models and the global model, while Huang et al. [124] uses the sum of pairwise functions to capture the relations among the clients. The algorithms proposed by [101, 102, 164, 232] require the clients to maintain certain local states and update them in every round, such as the local dual variables [232] or the local models [101, 102, 164]. The algorithms developed in [124, 232] also require the server to store individual updates from the clients. Sattler et al. [223] also considered a multi-task learning framework, and proposed to use clustering in the federated setting. The algorithm developed in [223] requires the server to know the clients’ identifiers so that it can compute the pairwise similarities between the clients and then partition them into different clusters.

Another common approach, which can be viewed as a special case of the multi-task learning framework, is to split the entire model into two parts: a shared part that is jointly learned by all clients, and a local part that is personalized to each client. Examples include [13, 67, 103, 168], where their algorithms require that every client stores the local part of the model as the local state, which is updated every round. Agarwal et al. [3] considers an online-learning scenario and assumes that the client can observe new samples in every round. However, the algorithm proposed by Agarwal et al. [3] requires the server to store the local predictions from individual clients, which might violate the data privacy constraint required by many federated learning applications.

7.5.2 Algorithms that Do Not Require Client-side State or Identifier

A popular technique used in this category is **meta-learning**. The goal of meta-learning is to learn a learning algorithm from a variety of tasks, so that this learning algorithm can solve new tasks using a small number of training samples. In the federated setting, every client can be treated as a different task, and the goal is to meta-learn a learning algorithm that can generalize to unseen

⁷Note that not all algorithms developed under the multi-task learning framework fall into this category, as shown in Section 7.5.2.

clients. A typical objective function is as follows:

$$\min_{\mathbf{x}} \sum_{i=1}^M f_i(g_i(\mathbf{x})), \quad (23)$$

where $g_i(\cdot)$ represents a learning algorithm that produces the local model on the i -th client (i.e., the local personalization algorithm), and \mathbf{x} denotes the parameters that configure the learning algorithm. Note that meta-learning and **multi-task learning** are closely related. In fact, meta-learning is a popular approach for developing algorithms for solving multi-task learning problems.

Fallah et al. [78] proposed an algorithm under the MAML (model-agnostic meta-learning) framework [83], where $g_i(\mathbf{x}) = \mathbf{x} - \alpha \nabla f_i(\mathbf{x})$ denotes a learning algorithm that performs one step of gradient descent starting from \mathbf{x} . Exactly computing the gradient of the MAML objective is usually computationally inefficient, so first-order approximations such as Reptile [189] might be preferred. On the other hand, MAML may have lower sample complexity than Reptile and as a result works much better when the number of examples per task is low [7]. Ozkara et al. [195] proposed QuPeL, a *quantized* and *personalized* FL algorithm, that facilitates collective training with heterogeneous clients while respecting resource diversity. For personalization, they allow clients to learn *compressed personalized models* with *different* quantization parameters depending on their resources. The authors develop an alternating proximal gradient update for solving their quantized personalization problem and analyzed its convergence properties. QuPeL also outperforms the other competing methods in personalized learning. Khodak et al. [143] showed that the FEDAVG algorithm [181], followed by fine-tuning the global model on new client’s local data during the inference time, is closely related to the Reptile algorithm. Specifically, Khodak et al. [143] proposed an online meta-learning algorithm, which was later generalized by Li et al. [158] with privacy guarantees. Similar connections were later pointed out by Jiang et al. [129]. Despite being the simplest possible approach to personalization, local fine-tuning is shown to work well in large-scale datasets [290] as well as real-world on-device applications [264]. Besides local fine-tuning, another popular local adaptation algorithm is to interpolating between local and global models [70, 178, 227]. Splitting the entire model into a shared part and a local part is another natural approach to personalization. Unlike the stateful algorithms mentioned in Section 7.5.1, Singhal et al. [230] proposed an algorithm where the local part is reconstructed locally each time a client participates in a round, and showed that the proposed algorithm also fits in the meta-learning framework.

Besides meta-learning, **clustering** is also a popular approach to personalization. The idea is to cluster the clients and learn a single model for each cluster. Unlike the clustering algorithm [223] mentioned in Section 7.5.1, the algorithms developed by [87, 178] do not require client-side state or identifier, and they work as follows: in each round, the server sends all models (one for each cluster) to a set of clients; each client, after receiving the models, first identifies which cluster it belongs to and then computes updates for that specific model. Besides iterative methods, Dennis et al. [69] propose a one-shot federated clustering technique which leverages data heterogeneity to weaken the separation assumptions for clustering. It can be used as a light-weight pre-processing step with applications to personalization and client selection.

8 Concluding Remarks

Federated learning is an active and interdisciplinary research area, in which many challenging problems lie at the intersections of machine learning, optimization, privacy, security, cryptography, information theory, distributed systems and many other areas. Research on federated optimization centers around formulating and solving the optimization problems arising in federated learning settings. This manuscript is intended to provide concrete recommendations and guidance for practitioners

and researchers who are new to the field on how to design and evaluate federated optimization algorithms (see Sections 1 to 4). Moreover, this manuscript extensively discussed the practical considerations as constraints for designing federated optimization algorithms, and briefly touched system deployment in Section 5. A concise yet self-contained discussion of the basic theory analysis is provided in Section 6 to help readers get familiar with the theoretical concepts and gain insights. Some elementary introductions to the broader scope of FL research are also given to help researchers understand how federated optimization intersects with other important considerations in FL and why certain recommendations are made (see Section 7).

This manuscript places a greater emphasis on guidelines for the practical implementation of federated optimization methods than most of previous works that share a similar spirit. Federated optimization algorithms are often complicated because of the natural partitioning of server/client computation and the various constraints imposed by system design and privacy considerations. The evaluation of FL methods may require extra caution regarding train/validation/test set splits, metrics and hyperparameters. We extensively use generalized FedAvg-like algorithms to showcase the suggested evaluation principles, rather than directly comparing the performance of different algorithms to recommend a specific set of algorithms. These evaluation principles are not intended to be a comprehensive checklist for readers to strictly follow, but are used to inspire researchers with ideas for how to perform their own evaluations. Among the suggestions, we want to emphasize the importance of specifying the application settings and comparison under the same set of constraints. Understanding the scenarios where an algorithm can be applied and has advantages can be particularly interesting for practitioners. However, as deploying algorithms in a real-world FL system can be a privilege not accessible to many FL researchers, the experimental evaluations mostly depend on carefully designed simulations. This suggestion can be universal for the general machine learning research, but particularly important for federated learning where the simulation experiments and system deployment environments can be very different.

FL applications are often categorized as either cross-device or cross-silo settings. Much discussion in this manuscript is biased towards cross-device settings for several reasons: cross-device settings usually impose more strict constraints (like local computation, communication efficiency, client availability etc), which makes the practical consideration different from conventional distributed optimization; and, many of the authors are more familiar with the cross-device settings in practice. Particularly, the real-world system deployment discussed in Section 5 is primarily based on the Google FL system [39]. We hope that most of the suggestions can also be useful in cross-silo settings. However, cross-silo FL may impose a different set of system and privacy constraints that empower researchers to design sophisticated algorithms beyond the scope of this manuscript.

Previous federated optimization research extensively discussed communication constraints, data heterogeneity, and to some extent computational constraints. Though many open problems and challenges still exist in these areas, as discussed in Section 1.1, we believe privacy, security and system issues should be considered equally important factors. Privacy and system consideration often introduce constraints that are challenging for designing optimization algorithms, and a co-design of optimization and these factors can be important.

Optimization remains one of the key and active research areas in federated learning. Though this manuscript is probably too long to be considered concise, we hope the extended discussions provide clear pointers and won't interfere with the main messages. We sincerely hope that the guidelines discussed can be useful for both researchers and practitioners to design and apply their federated learning algorithms.

Acknowledgements and Notes

The authors thank the early feedback and review by Sean Augenstein, Kallista Bonawitz, Corinna Cortes, and Keith Rush. This work is supported by GCP credits provided by Google Cloud. The simulation experiments are implemented with the TensorFlow Federated package. This paper originated at the discussion session moderated by Zheng Xu and Gauri Joshi at the Workshop on Federated Learning and Analytics, virtually held June 29–30th, 2020. During the discussion, a general consensus about the need for a guide about federated optimization is reached.

Editors We thank the extra work of section editors for helping incorporating contributions, organizing contents and addressing concerns. This is not a complete list of contributors or section authors. The editors⁸ would be happy to field questions from interested readers, or connect readers with the authors most familiar with a particular section. Zheng Xu edited Sections 1 and 8 with the help of Jakub Konečný and Brendan McMahan; Gauri Joshi edited Section 2 with the help of Brendan McMahan, Jianyu Wang, and Zheng Xu; Jianyu Wang edited Section 3 with the help of Gauri Joshi, Zachary Charles, and Satyen Kale; Zachary Charles edited Section 4 with the help of Brendan McMahan and Sebastian Stich; Zachary Garrett edited Section 5 with the help of Zachary Charles and Zheng Xu; Manzil Zaheer edited Section 6 with the help of Jianyu Wang and Honglin Yuan; Peter Kairouz edited Section 7 with the help of Galen Andrew, Tian Li, Zachary Charles, Shanshan Wu, and Deepesh Data.

Experiments Zachary Charles, Jianyu Wang, and Zheng Xu designed and coordinated the simulation experiments in Section 4, and provided necessary infrastructure. Zachary Charles, Weikang Song and Shanshan Wu worked on the GLD-* MobileNetV2 experiments; Zhouyuan Huo, Zheng Xu and Chen Zhu worked on the StackOverflow transformer experiments; Advait Gadhikar, Luyang Liu and Jianyu Wang worked on the CIFAR-10 ResNet cross-silo experiments. Hubert Eichner, Katharine Daly, Brendan McMahan, Zachary Garrett, Weikang Song, and Zheng Xu designed and collected data for the basic model in Section 5.4.

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.
- [2] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=B7v4QMR6Z9w>.
- [3] Alekh Agarwal, John Langford, and Chen-Yu Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.
- [4] Naman Agarwal, Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. In *Advances in Neural Information Processing Systems*, pages 7564–7575, 2018.

⁸Zheng Xu {xuzheng@google.com} for Sections 1 and 8, Gauri Joshi {gaaurij@andrew.cmu.edu} for Section 2, Jianyu Wang {jianyuw1@andrew.cmu.edu} for Section 3; Zachary Charles {zachcharles@google.com} for Section 4; Zachary Garrett {zachgarrett@google.com} for Section 5, Manzil Zaheer {manzilzaheer@google.com} for Section 6, and Peter Kairouz {kairouz@google.com} for Section 7.

- [5] Saurabh Agarwal, Hongyi Wang, Shivaram Venkataraman, and Dimitris Papailiopoulos. On the utility of gradient compression in distributed training systems. *arXiv preprint arXiv:2103.00543*, 2021.
- [6] Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. In *International Conference on Learning Representations (ICLR)*, 2021.
- [7] Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. On data efficiency of meta-learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.
- [8] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1709–1720, 2017.
- [9] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Neural Information Processing Systems (NeurIPS)*, pages 4618–4628, 2018.
- [10] Dan Alistarh, Torsten Hoefer, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 5973–5983, 2018.
- [11] Zeyuan Allen-Zhu, Faeze Ebrahimi, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations (ICLR)*, 2021.
- [12] Apple. Designing for privacy (video and slide deck). Apple WWDC, <https://developer.apple.com/videos/play/wwdc2019/708>, 2019.
- [13] Manoj Ghuh, Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.
- [14] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. 2015.
- [15] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.
- [16] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 274–283, 2018.
- [17] Sean Augenstein, H Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, et al. Generative models for effective ml on private, decentralized datasets. *arXiv preprint arXiv:1911.06679*, 2019.
- [18] The TensorFlow Federated Authors. TensorFlow Federated Stack Overflow dataset, 2019. URL https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data.
- [19] The TensorFlow Federated Authors. TensorFlow Federated Google Landmark v2 dataset, 2020. URL https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/gldv2/load_data.

- [20] Eugene Bagdasaryan and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *CoRR*, abs/1905.12101, 2019. URL <http://arxiv.org/abs/1905.12101>.
- [21] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- [22] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, pages 638–667. Springer, 2019.
- [23] Borja Balle, Peter Kairouz, Brendan McMahan, Om Dipakbhai Thakkar, and Abhradeep Thakurta. Privacy amplification via random check-ins. In *Advances in Neural Information Processing Systems*, 2020.
- [24] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.
- [25] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *NeurIPS - Advances in Neural Information Processing Systems*, pages 8635–8645, 2019.
- [26] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.
- [27] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.
- [28] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.
- [29] Raef Bassily, Shay Moran, and Anupama Nandi. Learning from mixtures of private and public populations. *arXiv preprint arXiv:2008.00331*, 2020.
- [30] Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Neural Information Processing Systems (NeurIPS)*, pages 14668–14679, 2019.
- [31] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [32] James Henry Bell, Kallista A. Bonawitz, Adrià Gascón, Tancrede Lepoint, and Mariana Raykova. Secure single-server aggregation with (poly)logarithmic overhead. In *CCS*, pages 1253–1269. ACM, 2020.
- [33] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, pages 560–569. PMLR, 2018.
- [34] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Titouan Parcollet, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.
- [35] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

- [36] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning (ICML)*, 2019.
- [37] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Neural Information Processing Systems (NeurIPS)*, pages 119–129, 2017.
- [38] 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. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.
- [39] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé M Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In *MLSys 2019*, 2019. URL <https://arxiv.org/abs/1902.01046>.
- [40] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. *arXiv preprint arXiv:2004.11791*, 2020.
- [41] Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.
- [42] Shicong Cen, Huishuai Zhang, Yuejie Chi, Wei Chen, and Tie-Yan Liu. Convergence of distributed stochastic variance reduced methods without sampling extra data. *arXiv preprint arXiv:1905.12648*, 2019.
- [43] Zachary Charles and Jakub Konečný. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.
- [44] Zachary Charles and Jakub Konečný. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020.
- [45] Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyan, and Virginia Smith. On large-cohort training for federated learning. *arXiv preprint arXiv:2106.07820*, 2021.
- [46] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *arXiv preprint arXiv:2012.02409*, 2020.
- [47] Pratik Chaudhari, Carlo Baldassi, Riccardo Zecchina, Stefano Soatto, Ameet Talwalkar, and Adam Oberman. Parle: parallelizing stochastic gradient descent. *arXiv preprint arXiv:1707.00424*, 2017.
- [48] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.
- [49] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.
- [50] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous sgd. *arXiv preprint arXiv:1604.00981*, 2016.

- [51] Kai Chen and Qiang Huo. Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5880–5884. IEEE, 2016.
- [52] Lingjiao Chen, Hongyi Wang, Zachary B. Charles, and Dimitris S. Papailiopoulos. DRACO: Byzantine-resilient distributed training via redundant gradients. In *ICML*, pages 902–911, 2018.
- [53] Mingqing Chen, Rajiv Mathews, Tom Ouyang, and Françoise Beaufays. Federated learning of out-of-vocabulary words. *arXiv preprint arXiv:1903.10635*, 2019.
- [54] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 2020.
- [55] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016.
- [56] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. *arXiv preprint arXiv:2007.11707*, 2020.
- [57] Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.
- [58] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *POMACS*, 1(2):44:1–44:25, 2017.
- [59] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. Secureboost: A lossless federated learning framework. *arXiv preprint arXiv:1901.08755*, 2019.
- [60] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.
- [61] Yuval Dagan and Vitaly Feldman. Interaction is necessary for distributed learning with privacy or communication constraints. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 450–462, 2020.
- [62] Deepesh Data and Suhas N. Diggavi. Byzantine-tolerant distributed coordinate descent. In *International Symposium on Information Theory (ISIT)*, pages 2724–2728, 2019.
- [63] Deepesh Data and Suhas N. Diggavi. Byzantine-resilient SGD in high dimensions on heterogeneous data. In *International Symposium on Information Theory (ISIT)*, 2021. Available online on arXiv:2005.07866.
- [64] Deepesh Data and Suhas N. Diggavi. Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data. In *International Conference on Machine Learning (ICML)*, 2021.
- [65] Deepesh Data, Linqi Song, and Suhas N. Diggavi. Data encoding for byzantine-resilient distributed optimization. *IEEE Transactions on Information Theory*, 67(2):1117–1140, 2021.
- [66] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.
- [67] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

- [68] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 2020.
- [69] Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In *International Conference on Machine Learning*, 2021.
- [70] Canh T Dinh, Nguyen H Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.
- [71] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.
- [72] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.
- [73] Sanghamitra Dutta, Gauri Joshi, Soumyadip Ghosh, Parijat Dube, and Priya Nagpurkar. Slow and stale gradients can win the race: Error-runtime trade-offs in distributed sgd. In *International Conference on Artificial Intelligence and Statistics*, pages 803–812. PMLR, 2018.
- [74] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [75] Hubert Eichner, Tomer Koren, Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 1764–1773. PMLR, 2019.
- [76] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, and Abhradeep Thakurta. Encode, shuffle, analyze privacy revisited: formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.
- [77] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *International Conference on Knowledge Discovery and Data Mining*, 2004.
- [78] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.
- [79] Biyi Fang, Xiao Zeng, and Mi Zhang. NestDNN: Resource-Aware Multi-Tenant On-Device Deep Learning for Continuous Mobile Vision. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 115–127, New Delhi, India, 2018.
- [80] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Security Symposium*, 2020.
- [81] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.
- [82] Olivier Fercoq, Zheng Qu, Peter Richtárik, and Martin Takáč. Fast distributed coordinate descent for minimizing non-strongly convex losses. *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.
- [83] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017.

- [84] Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $O(1/n)$. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65. PMLR, 2017.
- [85] Badih Ghazi, Rasmus Pagh, and Ameya Velingker. Scalable and differentially private distributed aggregation in the shuffled model. *arXiv preprint arXiv:1906.08320*, 2019.
- [86] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.
- [87] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, 2020.
- [88] Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramchandran. Communication efficient and byzantine tolerant distributed learning. In *IEEE International Symposium on Information Theory, (ISIT)*, pages 2545–2550. IEEE, 2020.
- [89] Antonious M. Girgis, Deepesh Data, and Suhas Diggavi. Differentially private federated learning with shuffling and client self-sampling. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.
- [90] Antonious M. Girgis, Deepesh Data, Suhas N. Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In Arindam Banerjee and Kenji Fukumizu, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 2521–2529. PMLR, 2021. Journal version in *IEEE Journal on Selected Areas in Information Theory (JSAIT)*, 2(1): 464-478, 2021.
- [91] Google. Your chats stay private while messages improves suggestions. <https://support.google.com/messages/answer/9327902>, 2020.
- [92] Google. Federated analytics: Collaborative data science without data collection. Google AI Blog, <https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html>, 2020.
- [93] Google. Advancing health research with google health studies. <https://blog.google/technology/health/google-health-studies-app/>, 2020.
- [94] Google. Get personalized actions, app suggestions, and more with device personalization services. <https://support.google.com/pixelphone/answer/9565916>, 2020.
- [95] Google. Your voice and audio data stays private while google assistant improves. <https://support.google.com/assistant/answer/10176224>, 2021.
- [96] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. *arXiv preprint arXiv:2011.02828*, 2020.
- [97] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [98] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [99] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

- [100] Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized stochastic gradient langevin dynamics and hamiltonian monte carlo, 2021.
- [101] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.
- [102] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [103] Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint arXiv:2102.09743*, 2021.
- [104] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- [105] Andrew Hard, Kurt Partridge, Cameron Nguyen, Niranjana Subrahmanya, Aishanee Shah, Pai Zhu, Ignacio Lopez Moreno, and Rajiv Mathews. Training keyword spotting models on non-iid data with federated learning. *arXiv preprint arXiv:2005.10406*, 2020.
- [106] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- [107] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.
- [108] Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *The Journal of Machine Learning Research*, 18(1):3744–3780, 2017.
- [109] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.
- [110] Chaoyang He, Conghui Tan, Hanlin Tang, Shuang Qiu, and Ji Liu. Central server free federated learning over single-sided trust social networks. *arXiv preprint arXiv:1910.04956*, 2019.
- [111] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Fednas: Federated deep learning via neural architecture search. *arXiv preprint arXiv:2004.08546*, 2020.
- [112] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33, 2020.
- [113] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Xinghua Zhu, Jianzong Wang, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020. URL <https://fedml.ai>.
- [114] Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via resampling. In *SpicyFL - NeurIPS Workshop on Scalability, Privacy, and Security in Federated Learning*, 2020.

- [115] Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Secure byzantine-robust machine learning. In *SpicyFL - NeurIPS Workshop on Scalability, Privacy, and Security in Federated Learning*, 2020.
- [116] Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [117] Samuel Horváth, Chen-Yu Ho, Ľudovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.
- [118] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.
- [119] Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos I Venieris, and Nicholas D Lane. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *arXiv preprint arXiv:2102.13451*, 2021.
- [120] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning (ICML)*, pages 4387–4398. PMLR, 2020.
- [121] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.
- [122] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *European Conference on Computer Vision*, 2020.
- [123] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. FedMGDA+: Federated learning meets multi-objective optimization. *arXiv preprint arXiv:2006.11489*, 2020.
- [124] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. *arXiv preprint arXiv:2007.03797*, 2021.
- [125] Alex Ingerman and Krzys Ostrowski. *TensorFlow Federated*, 2019. URL <https://medium.com/tensorflow/introducing-tensorflow-federated-a4147aa20041>.
- [126] Intel and Consilient. Intel and consilient join forces to fight financial fraud with ai. <https://newsroom.intel.com/news/intel-consilient-join-forces-fight-financial-fraud-ai/>, December 2020.
- [127] Intel®. Intel® open federated learning, 2021. URL <https://github.com/intel/openfl>.
- [128] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning (ICML)*, pages 476–484, 2014.
- [129] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.
- [130] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning (ICML)*, pages 1724–1732. PMLR, 2017.

- [131] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.
- [132] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.
- [133] Peter Kairouz, Mónica Ribero, Keith Rush, and Abhradeep Thakurta. Dimension independence in unconstrained private erm via adaptive preconditioning. *arXiv preprint arXiv:2008.06570*, 2020.
- [134] Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. *International Conference on Machine Learning (ICML)*, 2021.
- [135] Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. *International Conference on Machine Learning (ICML)*, 2021.
- [136] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning (ICML)*, volume 97, pages 3252–3261, 2019.
- [137] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020.
- [138] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *International Conference on Machine Learning (ICML)*, 2020.
- [139] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning (ICML)*, 2021.
- [140] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
- [141] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.
- [142] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4519–4529, 2020.
- [143] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, 2019.
- [144] Mikhail Khodak, Renbo Tu, Tian Li, Liam Li, Maria-Florina Balcan, Virginia Smith, and Ameet Talwalkar. Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing. *arXiv preprint arXiv:2106.04502*, 2021.
- [145] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [146] Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, 2019.
- [147] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, 2020.
- [148] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.
- [149] Jakub Konečný, H. Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.
- [150] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Neural Information Processing Systems (NeurIPS)*, 2020.
- [151] Dmitry Kovalev, Anastasia Koloskova, Martin Jaggi, Peter Richtárik, and Sebastian U. Stich. A linearly convergent algorithm for decentralized optimization: Sending less bits for free! In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, 2021.
- [152] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [153] Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Device heterogeneity in federated learning: A superquantile approach. *arXiv preprint arXiv:2002.11223*, 2020.
- [154] Fan Lai, Yinwei Dai, Xiangfeng Zhu, and Mosharaf Chowdhury. Fedyscale: Benchmarking model and system performance of federated learning. *arXiv preprint arXiv:2105.11367*, 2021.
- [155] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.
- [156] Anusha Lalitha, Shubhanshu Shekhar, Tara Javidi, and Farinaz Koushanfar. Fully decentralized federated learning. In *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.
- [157] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.
- [158] Jeffrey Li, Mikhail Khodak, Sebastian Caldas, and Ameet Talwalkar. Differentially private meta-learning. In *International Conference on Learning Representations*, 2020.
- [159] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Feddane: A federated newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231. IEEE, 2019.
- [160] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.
- [161] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2020.

- [162] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.
- [163] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.
- [164] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 2021.
- [165] Wei Li and Andrew McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning (ICML)*, pages 577–584, 2006.
- [166] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.
- [167] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [168] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.
- [169] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.
- [170] Feng Liao, Hankz Hankui Zhuo, Xiaoling Huang, and Yu Zhang. Federated hierarchical hybrid networks for clickbait detection. *arXiv preprint arXiv:1906.00638*, 2019.
- [171] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2020.
- [172] Tao Lin, Sebastian U. Stich, and Martin Jaggi. Don’t use large mini-batches, use local SGD. In *International Conference on Learning Representations (ICLR)*, 2018.
- [173] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.
- [174] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations (ICLR)*, 2018.
- [175] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Computing*, 1(1):105–115, 2019.
- [176] David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.
- [177] Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtárik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning (ICML)*, pages 6692–6701. PMLR, 2020.

- [178] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.
- [179] Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 456–464. Association for Computational Linguistics, 2010.
- [180] Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018. URL <https://openreview.net/pdf?id=BJ0hF1Z0b>.
- [181] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017. Initial version posted on arXiv in February 2016.
- [182] MELLODDY. Melloddy project meets its year one objective: Deployment of the world’s first secure platform for multi-task federated learning in drug discovery among 10 pharmaceutical companies. <https://www.melloddy.eu/y1announcement>, September 2020.
- [183] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning (ICML)*, pages 3518–3527, 2018.
- [184] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning (ICML)*, 2019.
- [185] Philipp Moritz, Robert Nishihara, Ion Stoica, and Michael I Jordan. SparkNet: Training deep networks in spark. *arXiv preprint arXiv:1511.06051*, 2015.
- [186] Giorgi Nadiradze, Ilia Markov Shigang Li Amirmojtaba Sabour, Peter Davies, and Dan Alistarh. Decentralized SGD with asynchronous, local and quantized updates. *arXiv preprint arXiv:1910.12308*, 2019.
- [187] Angelia Nedić, Alex Olshevsky, and César A. Uribe. Distributed learning for cooperative inference, 2017.
- [188] Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1), 2013.
- [189] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.
- [190] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.
- [191] NVIDIA. Nvidia clara, 2019. URL <https://developer.nvidia.com/clara>.
- [192] NVIDIA. Triaging covid-19 patients: 20 hospitals in 20 days build ai model that predicts oxygen needs. <https://blogs.nvidia.com/blog/2020/10/05/federated-learning-covid-oxygen-needs/>, October 2020.
- [193] Seyed Ali Osia, Ali Shahin Shamsabadi, Sina Sajadmanesh, Ali Taheri, Kleomenis Katevas, Hamid R. Rabiee, Nicholas D. Lane, and Hamed Haddadi. A hybrid deep learning architecture for privacy-preserving mobile analytics, 2017.

- [194] Owkin. Story of the 1st federated learning model at owkin. <https://owkin.com/federated-learning/federated-model/>, 2020.
- [195] Kaan Ozkara, Navjot Singh, Deepesh Data, and Suhas N. Diggavi. Qupel: Quantized personalization with applications to federated learning. *CoRR*, abs/2102.11786, 2021. URL <https://arxiv.org/abs/2102.11786>.
- [196] Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized langevin dynamics for bayesian learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15978–15989. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/b8043b9b976639acb17b035ab8963f18-Paper.pdf>.
- [197] Jihong Park, Sumudu Samarakoon, Mehdi Bennis, and Mérouane Debbah. Wireless network intelligence at the edge. *Proceedings of the IEEE*, 107(11):2204–2239, 2019.
- [198] Pitch Patarasuk and Xin Yuan. Bandwidth optimal all-reduce algorithms for clusters of workstations. *Journal of Parallel and Distributed Computing*, 69(2):117–124, 2009.
- [199] Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandeveld, et al. Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv preprint arXiv:2102.08503*, 2021.
- [200] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018.
- [201] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.
- [202] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455*, 2014.
- [203] Saurav Prakash, Sagar Dhakal, Mustafa Akdeniz, A. Salman Avestimehr, and Nageen Himayat. Coded computing for low-latency federated learning over wireless edge networks. *IEEE Journal on Selected Areas in Communication, Series on Machine Learning for Communications and Networks*, 2020.
- [204] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145 – 151, 1999.
- [205] Xun Qian, Hanze Dong, Peter Richtárik, and Tong Zhang. Error compensated loopless SVRG for distributed optimization. *OPT2020: 12th Annual Workshop on Optimization for Machine Learning (NeurIPS 2020 Workshop)*, 2020.
- [206] Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. *arXiv preprint arXiv:2010.00091*, 2020.
- [207] Shashank Rajput, Hongyi Wang, Zachary B. Charles, and Dimitris S. Papailiopoulos. DETOX: A redundancy-based framework for faster and more robust gradient aggregation. In *Neural Information Processing Systems (NeurIPS)*, pages 10320–10330, 2019.
- [208] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

- [209] S Reddi, Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Proceeding of 32nd Conference on Neural Information Processing Systems*, 2018.
- [210] Sashank J Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczós, and Alex Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.
- [211] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=LkFG31B13U5>.
- [212] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 2020.
- [213] Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity, 2020.
- [214] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016.
- [215] Jorma Rissanen and Glen G Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23(2):149–162, 1979.
- [216] Jae Hun Ro, Ananda Theertha Suresh, and Ke Wu. FedJAX: Federated learning simulation with JAX, 2020. URL <http://github.com/google/fedjax>.
- [217] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning (ICML)*, pages 8253–8265. PMLR, 2020.
- [218] Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, and Carlee Joe-Wong. Towards flexible device participation in federated learning for non-iid data. *arXiv preprint arXiv:2006.06954*, 2020.
- [219] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018.
- [220] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.
- [221] Sumudu Samarakoon, Mehdi Bennis, Walid Saad, and Merouane Debbah. Federated learning for ultra-reliable low-latency v2v communications. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–7. IEEE, 2018.
- [222] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [223] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

- [224] Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.
- [225] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [226] Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication efficient distributed optimization using an approximate Newton-type method. *arXiv preprint arXiv:1312.7853*, 2013.
- [227] Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Fei Wu, and Chao Wu. Federated mutual learning. *arXiv preprint arXiv:2006.16765*, 2020.
- [228] Navjot Singh, Deepesh Data, Jemin George, and Suhas N. Diggavi. Squarm-sgd: Communication-efficient momentum SGD for decentralized optimization. In *IEEE International Symposium on Information Theory (ISIT)*. arXiv preprint arXiv:2005.07041.
- [229] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33, 2020.
- [230] Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, Keith Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. *arXiv preprint arXiv:2102.03448*, 2021.
- [231] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.
- [232] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 2017.
- [233] Virginia Smith, Simone Forte, Ma Chenxin, Martin Takáč, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18:230, 2018.
- [234] Jinhyun So, Basak Guler, and A Salman Avestimehr. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning. *arXiv preprint arXiv:2002.04156*, 2020.
- [235] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.
- [236] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Neural Information Processing Systems (NeurIPS)*, pages 3517–3529, 2017.
- [237] Sebastian U Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.
- [238] Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.
- [239] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 4447–4458, 2018.

- [240] Nikko Strom. Scalable distributed dnn training using commodity gpu cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [241] Hang Su and Haoyu Chen. Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*, 2015.
- [242] Lili Su and Jiaming Xu. Securing distributed gradient descent in high dimensional statistical learning. *POMACS*, 3(1):12:1–12:41, 2019.
- [243] Gan Sun, Yang Cong, Jiahua Dong, Qiang Wang, and Ji Liu. Data poisoning attacks on federated machine learning. *arXiv preprint arXiv:2004.10020*, 2020.
- [244] Ziteng Sun, Peter Kairouz, A. T. Suresh, and H. McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.
- [245] Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *International Conference on Machine Learning (ICML)*, pages 3329–3337. PMLR, 2017.
- [246] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1139–1147. PMLR, 2013.
- [247] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning (ICML)*, 2019.
- [248] TFF. Tensorflow federated, 2020. URL <https://www.tensorflow.org/federated>.
- [249] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9): 803–812, 1986. doi: 10.1109/TAC.1986.1104412.
- [250] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [251] Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P Cunningham, David Schiminovich, and Christian P Robert. Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21(17):1–53, 2020.
- [252] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018.
- [253] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [254] Aidmar Wainakh, Alejandro Sanchez Guinea, Tim Grube, and Max Mühlhäuser. Enhancing privacy via hierarchical federated learning. *arXiv preprint arXiv:2004.11361*, 2020.
- [255] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in Neural Information Processing Systems (NeurIPS)*, 31:9850–9861, 2018.

- [256] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [257] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020.
- [258] Hongyi Wang, Saurabh Agarwal, and Dimitris Papailiopoulos. Pufferfish: Communication-efficient models at no extra cost. *Proceedings of Machine Learning and Systems*, 3, 2021.
- [259] Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd. *arXiv preprint arXiv:1810.08313*, 2018.
- [260] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.
- [261] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [262] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. SlowMo: Improving communication-efficient distributed SGD with slow momentum. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SkxJ8REYPH>.
- [263] Jianyu Wang, Zheng Xu, Zachary Garrett, Zachary Charles, Luyang Liu, and Gauri Joshi. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021.
- [264] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.
- [265] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Eavesdrop the composition proportion of training labels in federated learning. *arXiv preprint arXiv:1910.06044*, 2019.
- [266] WeBank. Utilization of fate in anti money laundering through multiple banks. <https://www.fedai.org/cases/utilization-of-fate-in-anti-money-laundering-through-multiple-banks/>, 2020.
- [267] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*, 2020.
- [268] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *arXiv preprint arXiv:1705.07878*, 2017.
- [269] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2020.
- [270] Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local SGD for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020.

- [271] Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020.
- [272] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *International Conference on Machine Learning (ICML)*, 2018.
- [273] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- [274] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.
- [275] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. SLSGD: secure and efficient distributed on-device machine learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD, Proceedings, Part II*, pages 213–228, 2019.
- [276] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning (ICML)*, pages 6893–6901, 2019.
- [277] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of Empires: Breaking Byzantine-tolerant SGD by Inner Product Manipulation. In *Uncertainty in Artificial Intelligence Conference (UAI)*, volume 115, pages 261–270. PMLR, 2020.
- [278] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno++: Robust fully asynchronous sgd. In *International Conference on Machine Learning (ICML)*, 2020.
- [279] Cong Xie, Sanmi Koyejo, Indranil Gupta, and Haibin Lin. Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. In *OPT workshop, NeurIPS*, 2020.
- [280] Cong Xie, Shuai Zheng, Oluwasanmi O Koyejo, Indranil Gupta, Mu Li, and Haibin Lin. Cser: Communication-efficient SGD with error reset. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.
- [281] Hang Xu, Chen-Yu Ho, Ahmed M Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos Karatsenidis, Marco Canini, and Panos Kalnis. Compressed communication for distributed deep learning: Survey and quantitative evaluation. Technical report, 2020.
- [282] Chengxu Yang, Wang Qipeng, Mengwei Xu, Zhenpeng Chen, Bian Kaigui, Yunxin Liu, and Xuanzhe Liu. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *The World Wide Web Conference*, 2021.
- [283] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. 2021.
- [284] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–19, 2019.
- [285] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions, 2018. URL <https://arxiv.org/abs/1812.02903>.

- [286] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, pages 5636–5645, 2018.
- [287] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Defending against saddle point attack in byzantine-robust distributed learning. In *ICML*, pages 7074–7084, 2019.
- [288] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning (ICML)*, pages 7184–7193. PMLR, 2019.
- [289] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.
- [290] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.
- [291] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [292] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. *arXiv preprint arXiv:2011.08474*, 2020.
- [293] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.
- [294] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019.
- [295] Chi Zhang and Qianxiao Li. Distributed optimization for over-parameterized learning. *arXiv preprint arXiv:1906.06205*, 2019.
- [296] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.
- [297] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*, 2020.
- [298] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33, 2020.
- [299] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.
- [300] Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning (ICML)*, pages 362–370, 2015.
- [301] Ligeng Zhu and Song Han. Deep leakage from gradients. In *Federated Learning*, pages 17–31. Springer, 2020.
- [302] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

A Datasets

A.1 GLD-23k and GLD-160k Datasets

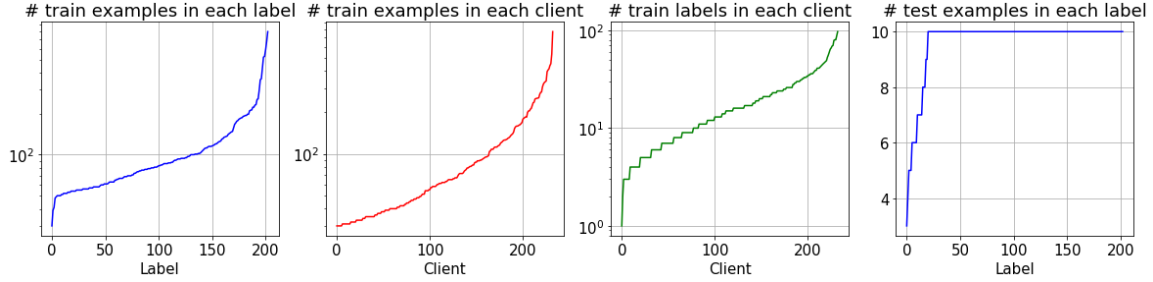


Figure 10: Statistics of the GLD-23k dataset.

The GLD-160k dataset is proposed by Hsu et al. [122] (under the name Landmarks-User-160k). It is a federated version of the centralized Google Landmark Dataset [190]. Each client represents an author, and the images are partitioned into clients based on their original authorship. GLD-160k contains 164,172 training images, 2028 landmark labels, and 1,262 clients. It has a single test set (i.e., the test dataset is not federated) containing 19,526 images. The GLD-23k dataset is a subset of the GLD-160k dataset. GLD-23k contains 23,080 training images, 203 landmark labels, and 233 clients. Same as GLD-160k, GLD-23k has a single test set containing 1,959 test examples, which covers all the labels in the federated training set (see the rightmost plot of Figure 10). Though smaller, the GLD-23k dataset maintains the imbalanced training characteristics of the GLD-160k dataset [122]. As shown in the left three plots of Figure 10, the number of training examples per label and the number of training examples per client both can vary for an order of magnitude, and the number of training labels in each client can vary for up to two orders of magnitude. These characteristics make the GLD-23k dataset representative of some issues endemic to real-world federated learning problems, and the smaller size makes it easier to work with in simulations.

A.2 Stack Overflow Dataset

Stack Overflow is a language dataset consisting of question and answers from the Stack Overflow online forum. The questions and answers have associated metadata, including tags (e.g., “javascript”), the creation time of the post, the title of the associated question, the score of the question, and the type of post (i.e., whether it is a question or an answer). Each client corresponds to a user, whose examples are all of their posts. We use the version from [18], which partitions the dataset among training, validation, and test clients. There are 342,477 train clients, 38,758 validation clients, and 204,088 test clients. Notably, the train clients only have examples from before 2018-01-01 UTC, while the test clients only have examples from after 2018-01-01 UTC. The validation clients have examples with no date restrictions, and all validation examples are held-out from both the test and train sets. The dataset is relatively unbalanced, with some users having very few posts, and other having many more (one client has over 80,000 examples).

A.3 CIFAR-10 Dataset

The CIFAR-10 dataset [152] is a computer vision dataset consisting of $32 \times 32 \times 3$ images with 10 possible labels. There are 50,000 training examples and 10,000 test examples, for a total of 6000

images per label. While this dataset does not have a natural partition among clients, Hsu et al. [121] propose a technique for creating non-IID partitions of the dataset across clients. Each client draws a multinomial distribution over the 10 labels from an underlying symmetric Dirichlet distribution with parameter α . To assign the client examples, we draw labels from this multinomial distribution, and sample (without replacement) one of the images in CIFAR-10 with the corresponding label. We proceed until the entire dataset has been exhausted. As discussed by Hsu et al. [121], this recovers a purely IID split when $\alpha \rightarrow \infty$, while for $\alpha = 0$ each client draws examples with a single label. We apply this technique with a Dirichlet parameter of $\alpha = 1$ to partition the training set of CIFAR-10 across 10 clients, each with 5000 images.

B Empirical Evaluation - Details

B.1 Algorithms

In Section 4, we refer to three algorithms, Algorithm A, B, and C. The purpose of this name obfuscation was to avoid making comparisons between algorithmic efficacy, as the purpose of that section was to provide suggestions for *how* to evaluate algorithms. For the interested reader, these algorithms are the following:

- Algorithm A: FEDAVG [181]
- Algorithm B: FEDAVGM [121]
- Algorithm C: FEDADAM [211]

All three algorithms are special cases of the generalized FEDOPT framework proposed by Reddi et al. [211] and presented in Algorithm 1. All three algorithms use SGD as the client optimizer (CLIENTOPT in Algorithm 1), but use different server optimizers (SERVEROPT in Algorithm 1). Specifically, FEDAVG, FEDAVGM, and FEDADAM use SGD, SGD with momentum, and ADAM [145]. Note that all three methods therefore perform the same amount of client computation and communication at each round.

In our experimental implementations, we only vary the client learning rate η and server learning rate η_s . For FEDAVGM, we use a momentum parameter of 0.9. For FEDADAM, we use server ADAM with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 0.001$. These are used for first-moment momentum, second-moment momentum, and numerical stability, respectively (see [145] for more details).

B.2 Models

In this section, we give a brief summary of the models used for each task in Section 4. We use one model per dataset.

CIFAR-10 For CIFAR-10 experiments, we train a modified ResNet-18 model, where the batch normalization layers are replaced by group normalization layers [273]. We use two groups in each group normalization layer. This replacement of batch normalization by group normalization was first proposed by Hsieh et al. [120], who show that it can lead to significant gains in heterogeneous federated settings.

GLD-23k and GLD-160k For both GLD-23k and GLD-160k experiments, we use a modified version of the MobileNetV2 model [222]. As with CIFAR-10, we replace all batch normalization layers with group normalization layers [273]. Each group normalization layer has two groups. We also do not use the dropout layers (setting their dropout probability to 0).

Stack Overflow For Stack Overflow experiments, we train a modified 3-layer Transformer model [250], where the dimension of the token embeddings is 96, and the hidden dimension of the feed-forward network (FFN) block is 1536. We use 8 heads for the multi-head attention, where each head is based on 12-dimensional (query, key, value) vectors. We use ReLU activation and set dropout rate to 0.1.

C Additional Experimental Results

C.1 More results for Section 4.2

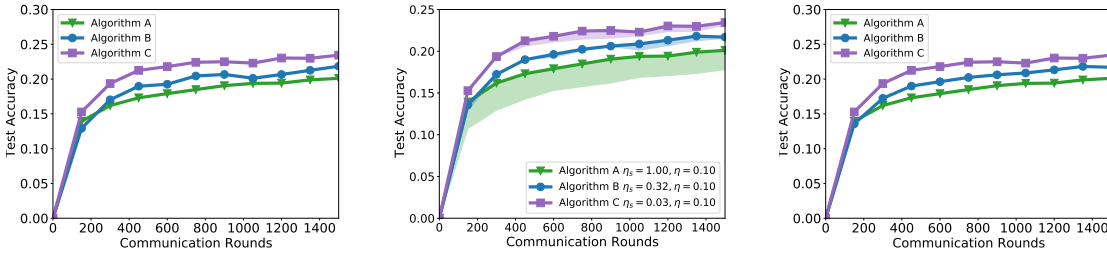


Figure 11: Test accuracy on Stack Overflow for Algorithms A, B, and C. The best performing client and server learning rate combinations are selected for each algorithm based on (a) training (b) validation (also see Figure 1) (c) testing accuracy. The trend of the curves are similar. For large-scale cross-device dataset, selecting by training and selecting by validation can be closely related because the probability of a client being sampled multiple times can be low. Selecting by testing can show the capacity (upper bound) of an algorithm, but is not a practical tuning strategy. More discussion in Section 4.2.1.

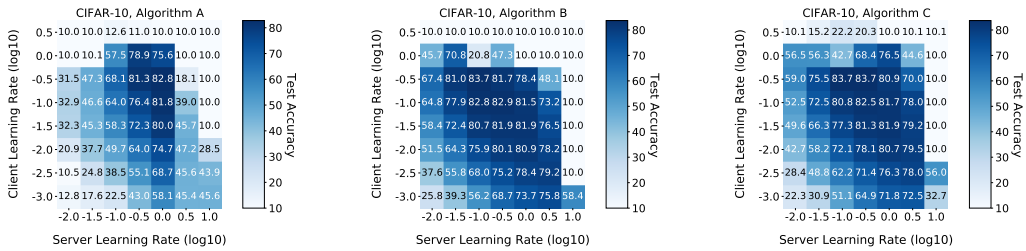


Figure 12: Test accuracy on CIFAR10 for various client and server learning rates. Results for Algorithms A, B, and C are given in the left, middle, and right plots, respectively. The test accuracy can show the hyperparameter sensitivity, while is unpractical for tuning.

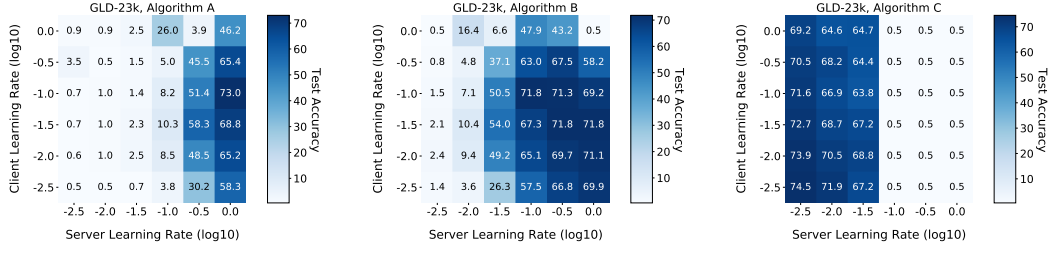


Figure 13: Test accuracy on GLD23k for various client and server learning rates. Results for Algorithms A, B, and C are given in the left, middle, and right plots, respectively. The test accuracy can show the hyperparameter sensitivity, while is unpractical for tuning.

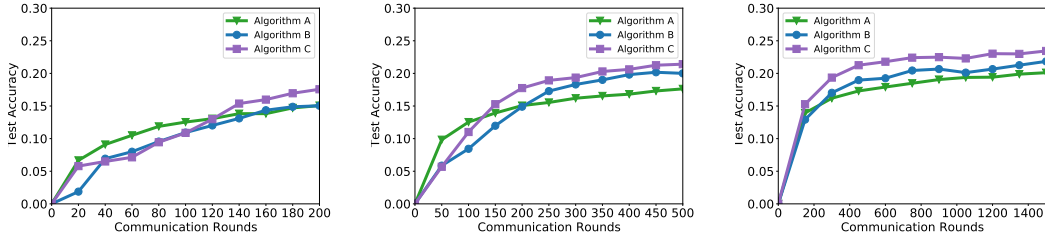


Figure 14: Validation accuracy on StackOverflow for a total of 200, 500, and 1000 communication rounds (left, middle, and right, respectively).

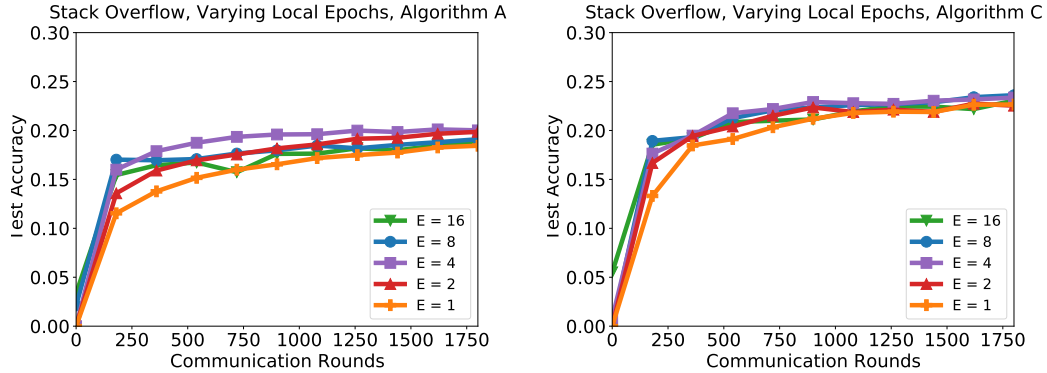


Figure 15: Validation accuracy on StackOverflow for Algorithms A (left) and C (right) for various numbers of local epochs per round E , versus the number of communication rounds.

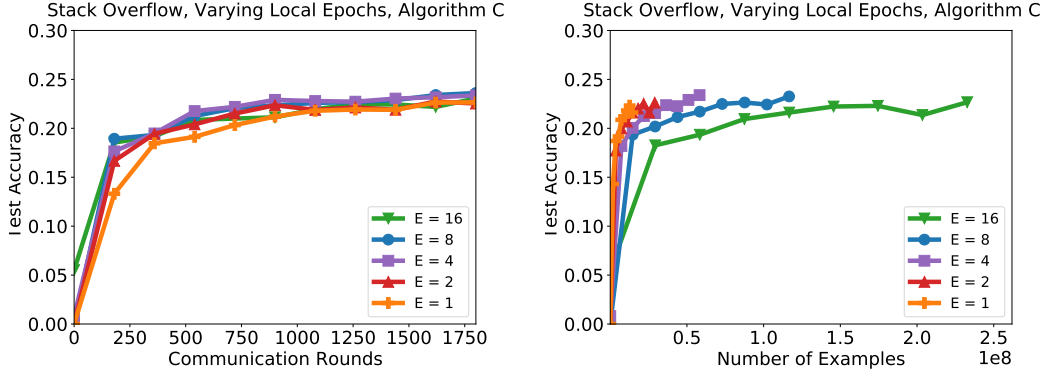


Figure 16: Validation accuracy on StackOverflow for Algorithms C. We plot validation accuracy versus the number of communication rounds (left) and versus the total number of examples processed by clients (right) for various numbers of local epochs per round E . We use learning rates $\eta = 0.01, \eta_s = 1.0$.

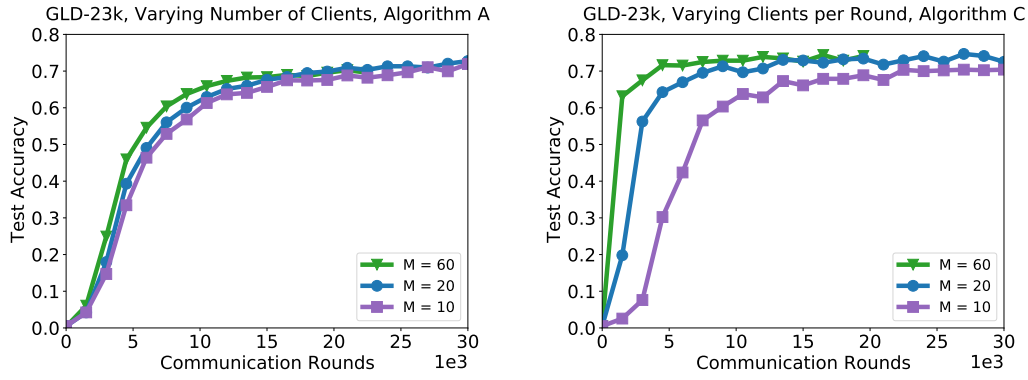


Figure 17: Test accuracy on GLD-23k for Algorithms A (left) and C (right) for various cohort sizes M , versus the number of communication rounds.

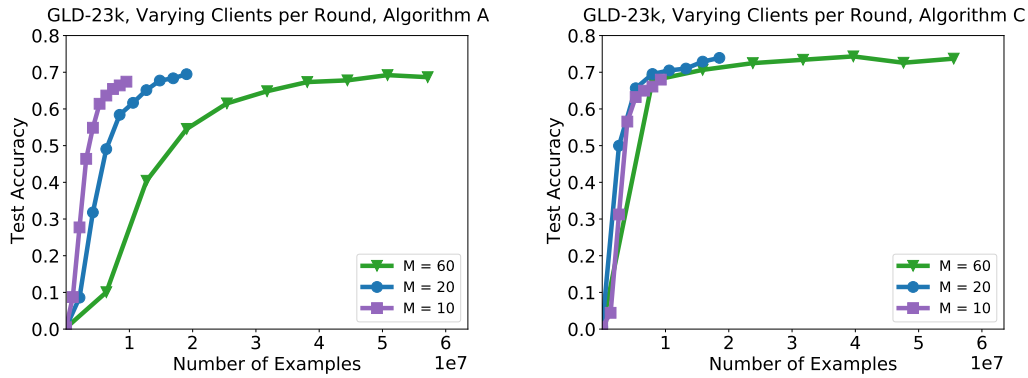


Figure 18: Test accuracy on GLD-23k for Algorithms A (left) and C (right) for various cohort size M , versus the number of examples processed by the clients.

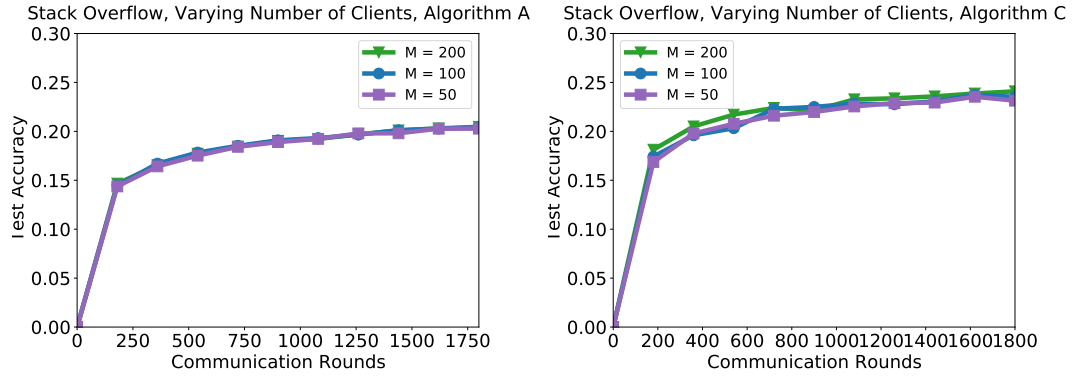


Figure 19: Validation accuracy on StackOverflow for Algorithms A (left) and C (right) for various cohort size M , versus the number of communication rounds.

C.2 Basic Model to Estimate On-Device Training Times in Section 5.4

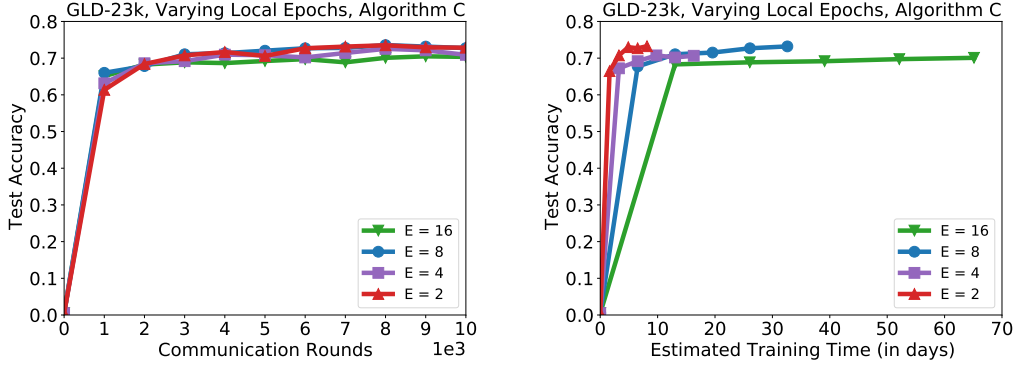


Figure 20: Test accuracy on GLD-23k for Algorithm C, for varying numbers of local epochs per round. We plot the test accuracy versus the number of communication rounds (left) and the estimate completion time using the model above (right).

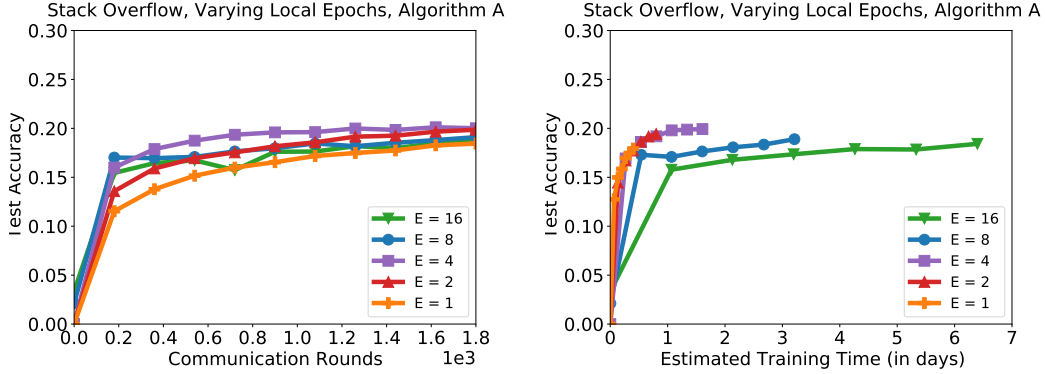


Figure 21: Accuracy of Algorithm A on a fixed set of 10,000 randomly selected test examples, for varying numbers of local epochs per round E . We plot the test accuracy versus the number of communication rounds (left) and the estimate completion time using the model above (right).

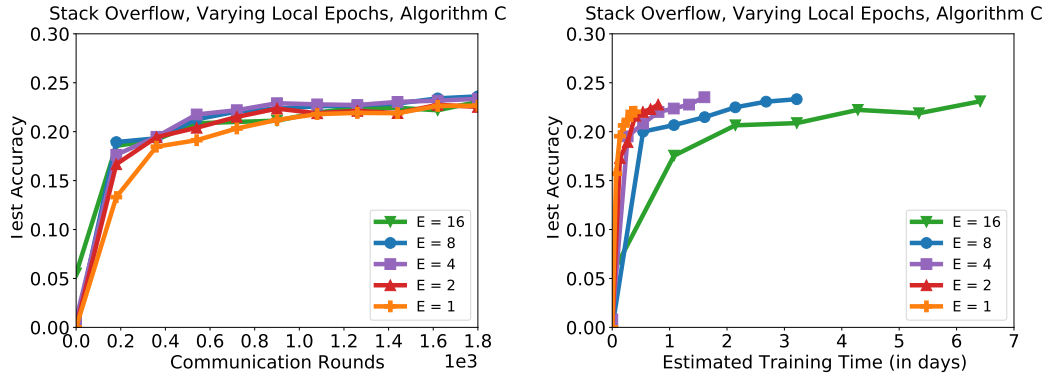


Figure 22: Accuracy of Algorithm C on a fixed set of 10,000 randomly selected test examples, for varying numbers of local epochs per round E . We plot the test accuracy versus the number of communication rounds (left) and the estimate completion time using the model above (right).

D Proofs

D.1 Deferred Proof of Lemma 1

Proof of Lemma 1. Since $\bar{\mathbf{x}}^{(t,k+1)} = \bar{\mathbf{x}}^{(t,k)} - \eta \frac{1}{M} \sum_{i=1}^M g_i(\mathbf{x}_i^{(t,k)})$, by parallelogram law

$$\frac{1}{M} \sum_{i=1}^M \left\langle g_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\rangle = \frac{1}{2\eta} \left(\left\| \bar{\mathbf{x}}^{(t,k)} - \mathbf{x}^* \right\|^2 - \left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 - \left\| \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\|^2 \right). \quad (24)$$

By convexity and L -smoothness of F_i , one has

$$\begin{aligned} F_i(\bar{\mathbf{x}}^{(t,k+1)}) &\leq F_i(\mathbf{x}_i^{(t,k)}) + \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)} \right\rangle + \frac{L}{2} \left\| \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)} \right\|^2 \\ &\quad (L\text{-smoothness}) \\ &\leq F_i(\mathbf{x}^*) + \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)} \right\|^2 \\ &\quad (\text{convexity}) \\ &\leq F_i(\mathbf{x}^*) + \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\rangle + L \left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 + L \left\| \mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \end{aligned} \quad (25)$$

Combining (24) and (25) yields

$$\begin{aligned} F(\bar{\mathbf{x}}^{(t,k+1)}) - F(\mathbf{x}^*) &= \frac{1}{M} \sum_{i=1}^M \left(F_i(\bar{\mathbf{x}}^{(t,k+1)}) - F(\mathbf{x}^*) \right) \\ &\leq \frac{1}{M} \sum_{i=1}^M \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}) - g_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\rangle + L \left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 + \frac{L}{M} \sum_{i=1}^M \left\| \mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \\ &\quad + \frac{1}{2\eta} \left(\left\| \bar{\mathbf{x}}^{(t,k)} - \mathbf{x}^* \right\|^2 - \left\| \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\|^2 - \left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \right). \end{aligned} \quad (26)$$

Since $\mathbb{E} \left[\nabla F_i(\mathbf{x}_i^{(t,k)}) - g_i(\mathbf{x}_i^{(t,k)}) \middle| \mathcal{F}^{(t,k)} \right] = 0$ we have

$$\begin{aligned} &\mathbb{E} \left[\frac{1}{M} \sum_{i=1}^M \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}) - g_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\rangle \middle| \mathcal{F}^{(t,k)} \right] \\ &= \mathbb{E} \left[\frac{1}{M} \sum_{i=1}^M \left\langle \nabla F_i(\mathbf{x}_i^{(t,k)}) - g_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\rangle \middle| \mathcal{F}^{(t,k)} \right] \\ &\leq \eta \cdot \mathbb{E} \left[\left\| \frac{1}{M} \sum_{i=1}^M (\nabla F_i(\mathbf{x}_i^{(t,k)}) - g_i(\mathbf{x}_i^{(t,k)})) \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] + \frac{1}{4\eta} \cdot \mathbb{E} \left[\left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] \\ &\quad (\text{Young's inequality}) \\ &\leq \frac{\eta \sigma^2}{M} + \frac{1}{4\eta} \cdot \mathbb{E} \left[\left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \middle| \mathcal{F}^{(t,k)} \right], \end{aligned} \quad (27)$$

where the last inequality is by bounded covariance assumptions and independence across clients. Plugging (27) back to the conditional expectation of (26) and noting that $\eta \leq \frac{1}{4L}$ yield

$$\begin{aligned}
& \mathbb{E} \left[F(\bar{\mathbf{x}}^{(t,k+1)}) - F(\mathbf{x}^*) \middle| \mathcal{F}^{(t,k)} \right] + \frac{1}{2\eta} \left(\mathbb{E} \left[\left\| \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}^* \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] - \left\| \bar{\mathbf{x}}^{(t,k)} - \mathbf{x}^* \right\|^2 \right) \\
& \leq \frac{\eta\sigma^2}{M} - \left(\frac{1}{4\eta} - L \right) \mathbb{E} \left[\left\| \bar{\mathbf{x}}^{(t,k+1)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] + \frac{L}{M} \sum_{i=1}^M \left\| \mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \\
& \leq \frac{\eta\sigma^2}{M} + \frac{L}{M} \sum_{i=1}^M \left\| \mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2. \quad (\text{since } \eta \leq \frac{1}{4L})
\end{aligned}$$

Telescoping k from 0 to τ completes the proof of Lemma 1. \square

D.2 Deferred Proof of Lemma 2

Proof of Lemma 2.

$$\begin{aligned}
& \mathbb{E} \left[\left\| \mathbf{x}_1^{(t,k+1)} - \mathbf{x}_2^{(t,k+1)} \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] = \mathbb{E} \left[\left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} - \eta \left(g_1(\mathbf{x}_1^{(t,k)}) - g_2(\mathbf{x}_2^{(t,k)}) \right) \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] \\
& \leq \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\|^2 - 2\eta \left\langle \nabla F_1(\mathbf{x}_1^{(t,k)}) - \nabla F_2(\mathbf{x}_2^{(t,k)}), \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\rangle \\
& \quad + \eta^2 \left\| \nabla F_1(\mathbf{x}_1^{(t,k)}) - \nabla F_2(\mathbf{x}_2^{(t,k)}) \right\|^2 + 2\eta^2 \sigma^2
\end{aligned}$$

Since $\max_i \sup_{\mathbf{x}} \left\| \nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x}) \right\| \leq \zeta$, the second term is bounded as

$$\begin{aligned}
& - \left\langle \nabla F_1(\mathbf{x}_1^{(t,k)}) - \nabla F_2(\mathbf{x}_2^{(t,k)}), \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\rangle \\
& \leq - \left\langle \nabla F(\mathbf{x}_1^{(t,k)}) - \nabla F(\mathbf{x}_2^{(t,k)}), \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\rangle + 2\zeta \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\| \\
& \leq - \frac{1}{L} \left\| \nabla F(\mathbf{x}_1^{(t,k)}) - \nabla F(\mathbf{x}_2^{(t,k)}) \right\|^2 + 2\zeta \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\| \quad (\text{by smoothness and convexity}) \\
& \leq - \frac{1}{L} \left\| \nabla F(\mathbf{x}_1^{(t,k)}) - \nabla F(\mathbf{x}_2^{(t,k)}) \right\|^2 + \frac{1}{2\eta\tau} \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\|^2 + 2\eta\tau\zeta^2 \quad (\text{by AM-GM inequality})
\end{aligned}$$

Similarly the third term is bounded as

$$\left\| \nabla F_1(\mathbf{x}_1^{(t,k)}) - \nabla F_2(\mathbf{x}_2^{(t,k)}) \right\|^2 \leq 3 \left\| \nabla F(\mathbf{x}_1^{(t,k)}) - \nabla F(\mathbf{x}_2^{(t,k)}) \right\|^2 + 6\zeta^2. \quad (28)$$

Replacing the above two bounds back to (by smoothness and convexity) gives (note that $\eta \leq \frac{1}{4L}$)

$$\begin{aligned}
\mathbb{E} \left[\left\| \mathbf{x}_1^{(t,k+1)} - \mathbf{x}_2^{(t,k+1)} \right\|^2 \middle| \mathcal{F}^{(t,k)} \right] & \leq \left(1 + \frac{1}{\tau} \right) \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\|^2 + 4\tau\eta^2\zeta^2 + 6\eta^2\zeta^2 + 2\eta^2\sigma^2 \\
& \leq \left(1 + \frac{1}{\tau} \right) \left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\|^2 + 10\tau\eta^2\zeta^2 + 2\eta^2\sigma^2.
\end{aligned}$$

Telescoping

$$\mathbb{E} \left[\left\| \mathbf{x}_1^{(t,k)} - \mathbf{x}_2^{(t,k)} \right\|^2 \middle| \mathcal{F}^{(t,0)} \right] \leq \frac{\left(1 + \frac{1}{\tau} \right)^k - 1}{\frac{1}{\tau}} \cdot (10\tau\eta^2\zeta^2 + 2\eta^2\sigma^2) \leq 18\tau^2\eta^2\zeta^2 + 4\tau\eta^2\sigma^2. \quad (29)$$

By convexity, for any i ,

$$\mathbb{E} \left[\left\| \mathbf{x}_i^{(t,k)} - \bar{\mathbf{x}}^{(t,k)} \right\|^2 \middle| \mathcal{F}^{(t,0)} \right] \leq 18\tau^2\eta^2\zeta^2 + 4\tau\eta^2\sigma^2. \quad (30)$$

□