Sharp Bounds for Federated Averaging (Local SGD) and Continuous Perspective*

Margalit Glasgow†
Stanford University
mglasgow@stanford.edu

Honglin Yuan†
Stanford University
yuanhl@cs.stanford.edu

Tengyu Ma
Stanford University
tengyuma@stanford.edu

Abstract

Federated Averaging (FEDAVG), also known as Local SGD, is one of the most popular algorithms in Federated Learning (FL). Despite its simplicity and popularity, the convergence rate of FEDAVG has thus far been undetermined. Even under the simplest assumptions (convex, smooth, homogeneous, and bounded covariance), the best known upper and lower bounds do not match, and it is not clear whether the existing analysis captures the capacity of the algorithm. In this work, we first resolve this question by providing a lower bound for FEDAVG that matches the existing upper bound, which shows the existing FEDAVG upper bound analysis is not improvable. Additionally, we establish a lower bound in a heterogeneous setting that nearly matches the existing upper bound. While our lower bounds show the limitations of FEDAVG, under an additional assumption of third-order smoothness, we prove more optimistic state-of-the-art convergence results in both convex and non-convex settings. Our analysis stems from a notion we call *iterate bias*, which is defined by the deviation of the expectation of the SGD trajectory from the noiseless gradient descent trajectory with the same initialization. We prove novel sharp bounds on this quantity, and show intuitively how to analyze this quantity from a Stochastic Differential Equation (SDE) perspective.

1 Introduction

Federated Learning (FL) is an emerging distributed learning paradigm in which a massive number of clients collaboratively participate in the training process without disclosing their private local data to the public [Konecny et al., 2015]. Typically, federated learning is orchestrated by a central server who oversees the clients, e.g. mobile devices or a group of organizations. The training process combines local training of a model at the clients with infrequent aggregation of the locally trained models at the central server.

Reflecting the goal of minimizing a loss function aggregated across clients, we consider the distributed optimization problem $\min F(\mathbf{x}) := \frac{1}{M}\sum_{m=1}^{M} F_m(\mathbf{x})$, where each client $m \in [M]$ holds a local objective F_m realized by its local data distribution $\mathcal{D}_m$, namely $F_m(\mathbf{x}) := \mathbb{E}_{\xi\sim\mathcal{D}_m} f(\mathbf{x};\xi)$. Federated Learning is *heterogeneous* by design as $\mathcal{D}_m$ can vary across clients. In the special case when $\mathcal{D}_m \equiv \mathcal{D}$ for all clients m, the problem is called *homogeneous*.

Federated Averaging (FEDAVG, McMahan et al. 2017), also known as Local SGD (Stich 2019), is one of the most popular algorithms applied in Federated Learning. In its simplest form, FEDAVG proceeds in R communication rounds, where at the beginning of each round, a central server sends the current iterate to each of the M clients. Each client then locally takes K steps of SGD, and then returns its final iterate to the central server. The central server averages these iterates to obtain the first iterate of the next round. We state the FEDAVG algorithm formally in Algorithm 1.

*Please visit https://arxiv.org/abs/2111.03741 for the complete and latest version of this paper.
†Equal contribution.

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

Table 1: **Convergence Rates of FEDAVG**. Some lower order terms as $R \to \infty$ omitted. H: smoothness, R: number of rounds, K: local iterations per round, M: number of clients, σ: noise, $D : \|\mathbf{x}^{(0,0)} - \mathbf{x}^\star\|$. The lower and upper bound use a slightly different metric of heterogeneity (ζ and ζ_*), see Remark 3.2 for details. We bold the terms where our analysis improves upon previous work.

	Homogeneous (Assumption 1)	Heterogeneous (Assumption 1 and 3)
Previous Upper Bound	$\frac{HD^2}{KR} + \frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}$ [Khaled et al., 2020]	$\frac{HD^2}{KR} + \frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{H^{\frac{1}{3}}\zeta^{\frac{2}{3}}D^{\frac{4}{3}}}{R^{\frac{2}{3}}}$ [Khaled et al., 2020, Woodworth et al., 2020a]
Our Lower Bound	$\frac{\mathbf{HD^2}}{\mathbf{KR}} + \frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{\mathbf{K^{\frac{1}{3}}}R^{\frac{2}{3}}}$ Theorem 3.1	$\frac{HD^2}{KR} + \frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{\mathbf{K^{\frac{1}{3}}}R^{\frac{2}{3}}} + \frac{\mathbf{H^{\frac{1}{3}}}\zeta_*^{\mathbf{\frac{2}{3}}}\mathbf{D^{\frac{4}{3}}}}{\mathbf{R^{\frac{2}{3}}}}$ Theorem 3.3
Previous Lower Bound	$\frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{\mathbf{K^{\frac{2}{3}}}R^{\frac{2}{3}}}$ [Woodworth et al., 2020b]	$\frac{\sigma D}{\sqrt{MKR}} + \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{\mathbf{K^{\frac{2}{3}}}R^{\frac{2}{3}}} + \min\left(\frac{\mathbf{HD^2}}{\mathbf{R}}, \frac{\mathbf{H^{\frac{1}{3}}}\zeta_*^{\mathbf{\frac{2}{3}}}\mathbf{D^{\frac{4}{3}}}}{\mathbf{R^{\frac{2}{3}}}}\right)$ [Woodworth et al., 2020a]

Algorithm 1 Federated Averaging (FEDAVG)

procedure FEDAVG ($\mathbf{x}^{(0,0)}, \eta$)
for $r = 0, \ldots, R-1$ **do**
  on client for $m \in [M]$ **in parallel do**
    $\mathbf{x}_m^{(r,0)} \leftarrow \mathbf{x}^{(r,0)}$ ▷ broadcast current iterate
    for $k = 0, \ldots, K-1$ **do**
      $\xi_m^{(r,k)} \sim \mathcal{D}_m$
      $\mathbf{g}_m^{(r,k)} \leftarrow \nabla f(\mathbf{x}_m^{(r,k)}; \xi_m^{(r,k)})$
      $\mathbf{x}_m^{(r,k+1)} \leftarrow \mathbf{x}_m^{(r,k)} - \eta \cdot \mathbf{g}_m^{(r,k)}$ ▷ client update
  $\mathbf{x}^{(r+1,0)} \leftarrow \frac{1}{M}\sum_{m=1}^M \mathbf{x}_m^{(r,K)}$ ▷ server averaging

While the FEDAVG algorithm is popular in practice, a thorough theoretical understanding of FEDAVG has not been established. Even under the simplest setting (convex, smooth, homogeneous and bounded covariance, see Assumption 1), the state-of-the-art upper bounds for FEDAVG due to Khaled et al. [2020] and Woodworth et al. [2020b] do not match the state-of-the-art lower bound due to Woodworth et al. [2020b], see Table 1. This suggests that at least one side of the analysis is not sharp. Therefore a fundamental question remains:

Does the current convergence analysis of FEDAVG *fully capture the capacity of the algorithm?*

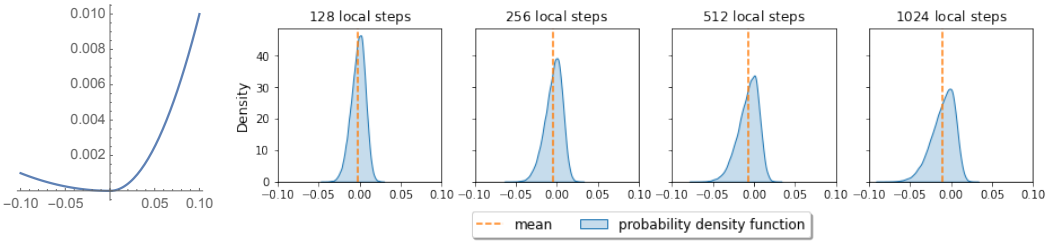
Our first contribution is to answer this question definitively under the standard smoothness and convexity assumptions. We establish a sharp lower bound for FEDAVG that matches the existing upper bound (Theorem 3.1), showing that the existing FEDAVG analysis is *not* improvable. Moreover, we establish a stronger lower bound in the *heterogeneous* setting, Theorem 3.3, which suggests the best known *heterogeneous* upper bound analysis [Khaled et al., 2020, Woodworth et al., 2020a] is also (almost)[3] not improvable.

Our proofs highlight exactly what can go wrong in FEDAVG, yielding these slow convergence rates. Specifically, our lower bound analysis stems from a notion we call *iterate bias*, which is defined by the deviation of the expectation of the SGD trajectory from the (noiseless) gradient descent trajectory with the same initialization (see Definition 2.1 for details). We show that even for convex and smooth objectives, the mean of SGD initialized at the optimum can drift away from the optimum at the rate of $\Theta(\eta^2 k^{\frac{3}{2}})$ after k steps,[4] for sufficiently small learning rate η. We depict this phenomenon in Fig. 1.[5]The iterate bias thus quantifies the fundamental difficulty encountered by FEDAVG:

[3]Up to a minor variation of the definition of heterogeneity measure, see Table 1.

[4]This rate is also sharp according to our matching upper and lower bounds, see Theorems 2.2 and 2.3 for details.

[5]Code see https://github.com/hongliny/Sharp-Bounds-for-FedAvg-and-Continuous-Perspective.



(a) Function plot

(b) Probability density function after various SGD steps

Figure 1: **Illustration of the iterate bias of SGD.** Consider the objective $F(x) = \begin{cases} x^2 & x \geq 0 \\ \frac{1}{10}x^2 & x < 0 \end{cases}$ as shown in (a), and $f(x;\xi) := \xi x + F(x)$ where $\xi \sim \mathcal{N}(0, 0.01)$. We initialize the SGD at optimum $x^\star = 0$, and run 1024 steps of SGD with step size 10^{-2}. We repeat this random process for 65536 times, and estimate the density function after 128, 256, 512 and 1024 steps. Observe that the density function and the average gradually move to the left (away from the optimum, where the curvature is smaller). This figure explains the intrinsic difficulty for FEDAVG to handle objective with drastic Hessian change.

Even with infinite number of homogeneous clients, FEDAVG *can drift away from the optimum even if initialized at the optimum.*

Indeed, we show in Section B.2 that the sharp lower bound of SGD iterate bias leads directly to our sharp lower bound of FEDAVG convergence rate.

The discouraging lower bound of FEDAVG under a standard smoothness assumption does not conform well with its empirical efficiency observed in practice [Lin et al., 2020c]. This motivates us to consider whether additional modeling assumptions could better explain the empirical performance of FEDAVG. The aforementioned lower bound is attained by a special piece-wise quadratic function with a sudden curvature change, which is smooth (has bounded second-order derivatives) but has unbounded third-order derivatives. A natural assumption to exclude this corner case is third-order smoothness, which has been considered before in the context of federated learning [Yuan and Ma, 2020], and may be representative of objectives in practice. For instance, loss functions used to learn many generalized linear models, such as logistic regression, often exhibit third-order smoothness [Hastie et al., 2009].

With this third-order smoothness assumption, we show that the iterate bias reduces to $\Theta(\eta^3 k^2)$, one order higher in η than the rate under only second-order smoothness.[6] While the proofs for bounding the iterate bias are quite technical, we show that it is easy to analyze the bias via a continuous approach. More specifically, by studying the stochastic differential equation (SDE) corresponding to the continuous limit of SGD, one can derive the limit of the iterate bias of generic objectives by using the Kolmogorov backward equation of the SDE, see Section 2.3.

Leveraging this intuition from the bias, we prove state-of-the-art rates for FEDAVG under third-order smoothness in *both* convex and non-convex settings (Theorems 4.1 and 4.2). In non-convex settings, our convergence rate scales with $1/R^{\frac{4}{5}}$, which improves upon the best known rate of $1/R^{\frac{2}{3}}$ [Yu et al., 2019b] if we do not assume third-order smoothness.

1.1 Organization and Notation

In Section 2, we formally define the iterate bias of SGD, and state sharp bounds on its rate. In Section 3, we state our lower bounds for FEDAVG, and show how the iterate bias can be used to achieve our sharp bounds. In Section 4, we state our convergence results for FEDAVG under third-order smoothness. All proofs are deferred to the appendix.

We use bold lower case character to denote vectors (e.g., $\mathbf{x}$). We use $\|\cdot\|$ to denote the ℓ_2-norm of a vector, $[n]$ to denote the set $\{1, \ldots, n\}$. Throughout the paper, we use O, Ω, Θ notation to hide absolute constants only.

We defer the literature review to Section A due to space constraints.

[6] This rate is sharp according to our matching upper and lower bounds, see Theorems 2.4 and 2.5.

2 Setup and Technical Overview: Intuition From Iterate Bias of SGD

The intuition from our lower bound comes from studying the behaviour of FEDAVG when the number of clients, M, tends to infinity. In this case, the averaged iterate $\mathbf{x}^{(r+1,0)}$ is precisely the *expected* iterate after K iterations of SGD starting from the last averaged iterate, $\mathbf{x}^{(r,0)}$. This motivates the following definition.

Definition 2.1 (Iterate Bias of SGD)**.** *Let $\{\mathbf{x}_{\texttt{SGD}}^{(k)}\}_{k=0}^{\infty}$ and $\{\mathbf{z}_{\texttt{GD}}^{(k)}\}_{k=0}^{\infty}$ be the trajectories of SGD and GD initialized at the same point $\mathbf{x}$, formally*

$$\mathbf{x}_{\texttt{SGD}}^{(k+1)} \leftarrow \mathbf{x}_{\texttt{SGD}}^{(k)} - \eta \nabla f(\mathbf{x}_{\texttt{SGD}}^{(k)}; \xi^{(k)}), \qquad \mathbf{x}_{\texttt{SGD}}^{(0)} = \mathbf{x};$$
$$\mathbf{z}_{\texttt{GD}}^{(k+1)} \leftarrow \mathbf{z}_{\texttt{GD}}^{(k)} - \eta \nabla F(\mathbf{z}_{\texttt{GD}}), \qquad \mathbf{z}_{\texttt{GD}}^{(0)} = \mathbf{x}.$$

*The **iterate bias** (or in short "bias") from $\mathbf{x}$ at the k-th step is defined as $\mathbb{E}\,\mathbf{x}_{\texttt{SGD}}^{(k)} - \mathbf{z}_{\texttt{GD}}^{(k)}$, the difference between the mean of SGD trajectory and the (deterministic) GD trajectory.*

One important special case of Definition 2.1 is the iterate bias from a stationary point $\mathbf{x}^\star$. In this case, the gradient descent trajectory $\mathbf{z}_{\texttt{GD}}^{(k)}$ will stay at the optimum since $\nabla F(\mathbf{z}_{\texttt{SGD}}^{(k)}) \equiv \nabla F(\mathbf{x}^\star) = \mathbf{0}$. The iterate bias then reduces to $\mathbb{E}[\mathbf{x}_{\texttt{SGD}}^{(k)}] - \mathbf{x}^\star$. Notably, even for convex smooth objectives f, the expected iterate $\mathbb{E}[\mathbf{x}_{\texttt{SGD}}^{(k)}]$ may drift away from the optimum $\mathbf{x}^\star$, even if initialized at the $\mathbf{x}^\star$. This occurs because of a difference between the gradient of the expectation of an iterate, $\nabla f(\mathbb{E}[\cdot])$, and the expectation of the gradient of the iterate, $\mathbb{E}[\nabla f(\cdot)]$.

In Fig. 1, we illustrate this phenomenon via a one-dimensional objective. This figure, and our formal results below, illustrate that for sufficiently small step sizes, the bias increases in k. For this reason, doing more than one local step can sometimes be counterproductive (when $k = 1$, the bias is always zero). This phenomenon is key to the poor dependence on K in the convergence rate we prove for FEDAVG.

2.1 The Bias Under Second-Order Smoothness

In this subsection, we provide sharp bounds on the iterate bias under standard assumptions, formally given below.

Assumption 1. *Assume $f(\mathbf{x}; \xi)$ is second-order differentiable w.r.t. $\mathbf{x}$, and*

(a) Convexity: $f(\mathbf{x}; \xi)$ is convex with respect to $\mathbf{x}$ for any ξ.

(b) Smoothness: $f(\mathbf{x}; \xi)$ is H-smooth with respect to $\mathbf{x}$. That is, for any ξ, for any $\mathbf{x}, \mathbf{y}$, we have $\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|_2 \le H\|\mathbf{x} - \mathbf{y}\|_2$.

(c) Bounded covariance: for any $\mathbf{x}$, $\mathbb{E}_{\xi \sim \mathcal{D}} \|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2 \le \sigma^2$.

We establish the following upper bound on the bias.[7]

Theorem 2.2 (Simplified from Theorem C.1)**.** *Under Assumption 1, there exists an absolute constant $\bar{c}$ such that for any initialization $\mathbf{x}$, for any $\eta \le \frac{1}{H}$, the iterate bias satisfies $\left\|\mathbb{E}\,\mathbf{x}_{\texttt{SGD}}^{(k)} - \mathbf{z}_{\texttt{GD}}^{(k)}\right\|_2 \le \bar{c} \cdot \eta^2 k^{\frac{3}{2}} H\sigma$.*

In fact, we show in the following theorem that this upper bound of iterate bias is sharp.

Theorem 2.3 (Simplified from Theorem C.2)**.** *There exists an absolute constant $\underline{c}$ such that for any H, σ, there exists an objective $f(\mathbf{x}; \xi)$ and distribution $\xi \sim \mathcal{D}$ satisfying Assumption 1 such that for any integer K, for any $\eta \le \frac{1}{2KH}$, and integer $k \in [2, K]$, the iterate bias from the optimum $\mathbf{x}^\star$ of F is lower bounded as $\left\|\mathbb{E}\,\mathbf{x}_{\texttt{SGD}}^{(k)} - \mathbf{z}_{\texttt{GD}}^{(k)}\right\|_2 \ge \underline{c} \cdot \eta^2 k^{\frac{3}{2}} H\sigma$.*

Theorem 2.3 shows that the SGD trajectory can indeed drift away (in expectation) from the optimum $\mathbf{x}^\star$ despite being initialized at $\mathbf{x}^\star$. Our lower bound improves over the best known lower bound

[7] Throughout this section, we mainly focus on the iterate bias bound in the regime of sufficiently small η for simplicity and easy comparison. Our complete theorem in appendix covers the case of general η choice.

$\Omega(\eta^2 k H\sigma)$ due to Woodworth et al. [2020b]. The lower bound is attained by running SGD with Gaussian noise on the piecewise quadratic function $f(x) := \begin{cases} \frac{1}{2}Hx^2 & x \geq 0, \\ \frac{1}{4}Hx^2 & x < 0. \end{cases}$, first analyzed in Woodworth et al. [2020b].

Recall that the bias originates from the difference between $\nabla f(\mathbb{E}[\mathbf{x}_{\texttt{SGD}}^{(k)}])$ and $\mathbb{E}[\nabla f(\mathbf{x}_{\texttt{SGD}}^{(k)})]$. This piecewise quadratic function has an unbounded third order derivative at 0, which causes this difference to be large whenever the distribution of $\mathbf{x}_{\texttt{SGD}}^{(k)}$ spans both sides of 0. This worst case construction motivates our further study of the bias under a third-order derivative bound.

2.2 The Bias Under Third-Order Smoothness

We formally state our third-order smoothness condition in the following assumption.

Assumption 2. *Assume $f(\mathbf{x};\xi)$ is third-order differentiable w.r.t. $\mathbf{x}$ for any ξ, and*

(a) $f(\mathbf{x};\xi)$ is Q-3rd-order-smooth, i.e. for any ξ, for any $\mathbf{x}, \mathbf{y}$, $\|\nabla^2 f(\mathbf{x};\xi) - \nabla^2 f(\mathbf{y};\xi)\|_2 \leq Q\|\mathbf{x}-\mathbf{y}\|_2$.

(b) $\nabla f(\mathbf{x},\xi)$ has σ^4-bounded 4th order central moment, i.e. for all $\mathbf{x}$, $\mathbb{E}_\xi\left[\|\nabla f(\mathbf{x},\xi) - \nabla F(\mathbf{x})\|^4\right] \leq \sigma^4$.

We show that under this additional assumption, the iterate bias reduces to $\mathrm{O}(\eta^3 k^2 Q\sigma^2)$, which scales on the order of η^3 (rather than η^2) as η goes to 0.

Theorem 2.4 (Simplified from Theorem C.3)**.** *Under Assumptions 1 and 2, there exists an absolute constant $\bar{c}$ such that for any initialization $\mathbf{x}$, for any $\eta \leq \frac{1}{2H}$, the iterate bias satisfies $\left\|\mathbb{E}\,\mathbf{x}_{\texttt{SGD}}^{(k)} - \mathbf{z}_{\texttt{GD}}^{(k)}\right\|_2 \leq \bar{c}\cdot\eta^3 k^2 Q\sigma^2$.*

Theorem 2.4 also reveals the dependency on the third-order smoothness Q. In the extreme case where $Q = 0$ (f is quadratic), the iterate bias will disappear. It is worth noting that since Assumption 1 is still required in Theorem 2.4, the original upper bound $\mathrm{O}(\eta^2 k^{\frac{3}{2}} H\sigma)$ from Theorem 2.2 still applies, and one can formulate the upper bound as the minimum of the two.

The following lower bound shows that the upper bound in Theorem 2.4 is sharp.

Theorem 2.5 (Simplified from Theorem C.4)**.** *There exists an absolute constant $\underline{c}$ such that for any H, σ, K, for any sufficiently small Q (polynomially dependent on H, σ, K), there exists an objective $f(\mathbf{x};\xi)$ and distribution $\xi \sim \mathcal{D}$ satisfying Assumptions 1 and 2 such that for any $\eta \leq \frac{1}{2HK}$ and integer $k \in [2, K]$, the iterate bias from the optimum $\mathbf{x}^\star$ is lower bounded as $\left\|\mathbb{E}\,\mathbf{x}_{\texttt{SGD}}^{(k)} - \mathbf{z}_{\texttt{GD}}^{(k)}\right\|_2 \geq \underline{c}\cdot\eta^3 k^2 Q\sigma^2$.*

2.3 Revealing Iterate Bias Via Continuous Perspective

While the proofs of the results above are quite technical, the intuition for these bounds is much easier to see in a continuous view of SGD. As an example, we demonstrate how the $\Theta(\eta^3 k^2 Q\sigma^2)$ term shows up in Theorems 2.4 and 2.5.

Consider a one-dimensional instance of SGD with Gaussian noise, where $f(x;\xi) = F(x) - \xi x$, and $\xi \sim \mathcal{N}(0,\sigma^2)$. The SGD then follows

$$x_{\texttt{SGD}}^{(k+1)} = x_{\texttt{SGD}}^{(k)} - \eta\nabla F(x_{\texttt{SGD}}^{(k)}) + \eta\xi^{(k)}, \quad \text{where } \xi^{(k)} \sim \mathcal{N}(0,\sigma^2). \tag{2.1}$$

The continuous limit of (2.1) corresponds to the following SDE, with the scaling $t = \eta k$:

$$\mathrm{d}X(t) = -F'(X(t))\mathrm{d}t + \sqrt{\eta}\sigma\mathrm{d}B_t, \tag{2.2}$$

where B_t denotes the Brownian motion (also known as the Wiener process).[8]

[8] To justify the relation of Eq. (2.1) and Eq. (2.2), note that Eq. (2.1) can be viewed as a numerical discretization (Euler-Maruyama discretization [Kloeden and Platen, 1992]) of the SDE (2.2) with time step-size η.

To get a handle of the iterate bias, our goal is to study $\mathbb{E}[X(t)|X(0)=x]$, the expectation of the SDE solution $X(t)$ initialized at x. We view this quantity as a multivariate function $u(t,x)$ of t and x, with the objective to Taylor expand $u(t,x)$ around $u(0,x)$ in t:

$$u(t,x) = u(0,x) + u_t(0,x)t + \frac{1}{2}u_{tt}(0,x)t^2 + o(t^2).$$

For brevity, we use subscript notation to denote partial derivatives, e.g, u_x denotes $\frac{\partial u(t,x)}{\partial x}$. The relationship of $u(t,x)$ and the SDE (2.2) is established by the Kolmogorov backward equation as follows.

Claim 2.6 (Kolmogorov backward equation [Øksendal, 2003])**.** *Let $u(t,x) = \mathbb{E}[X(t)|X(0)=x]$, then $u(t,x)$ satisfies the following partial differential equation:*

$$u_t = -F_x u_x + \eta\sigma^2 u_{xx}, \quad \text{with } u(0,x) = x. \tag{2.3}$$

Using this claim, we can compute the first two derivatives of $u(t,x)$ in t, as follows:

Lemma 2.7. *Suppose $u(t,x)$ satisfies the PDE (2.3), then $u_t(0,x) = -F_x, u_{tt}(0,x) = F_x F_{xx} - \eta\sigma^2 F_{xxx}$.*

With Lemma 2.7 we can expand $u(t,x)$ around $(0,x)$:

$$u(t,x) = x - F_x t + \frac{1}{2}\left(F_x F_{xx} - \eta\sigma^2 F_{xxx}\right)t^2 + o(t^2).$$

Ignoring higher order terms in t, the term $-\frac{1}{2}\eta\sigma^2 F_{xxx}$ reflects the difference between the noiseless GD trajectory from x and $\mathbb{E}[X(t)|X(0)=x]$, that is, the iterate bias. Converting back to the discrete trajectory (Eq. (2.1)) via the scaling $t = \eta k$, we obtain

$$\mathbb{E}[x_{\texttt{SGD}}^{(k)}] - z_{\texttt{GD}}^{(k)} \approx -\frac{1}{2}\eta^3 k^2 \sigma^2 F_{xxx}(x).$$

When the third derivative of F is bounded by Q, this recovers the upper bound of $O(\eta^3 k^2 Q\sigma^2)$ in Theorem 2.4. The lower bound of Theorem 2.5 follows by choosing a function with third derivative Q at $x^\star$.

3 Lower Bound Results

In this section, we present our lower bounds for FEDAVG in both convex homogeneous and heterogeneous settings, and discuss its implications. We then show how use the lower bound on the bias of SGD from Section 2 to establish a lower bound on the convergence of FEDAVG.

Our main result for the homogeneous setting is the following theorem.

Theorem 3.1 (Lower bound for homogeneous FEDAVG (see Theorem D.1))**.** *For any $K \geq 2$, R, M, σ, and D, there exists $f(\mathbf{x};\xi)$ and distribution $\xi \sim \mathcal{D}$ satisfying Assumption 1 with optimum $\mathbf{x}^\star$, such that for some initialization $\mathbf{x}^{(0,0)}$ with $\|\mathbf{x}^{(0,0)} - \mathbf{x}^\star\|_2 < D$, the final iterate of* FEDAVG *with any step size satisfies:*

$$\mathbb{E}\left[F(\mathbf{x}^{(R,0)})\right] - F(\mathbf{x}^\star) \geq \Omega\left(\frac{HD^2}{KR} + \frac{\sigma D}{\sqrt{MKR}} + \min\left\{\frac{\sigma D}{\sqrt{KR}}, \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right\}\right).$$

This lower bound matches the best upper bound given by the theorem 2 of [Woodworth et al., 2020b].

We extend our results to FEDAVG in the heterogeneous setting. Recall that in this setting, we allow each client m to draw ξ from its own distribution $\mathcal{D}_m$. We prove our results under the following assumption on heterogeneity of the gradient at the optimum.

Assumption 3 (Bounded gradient heterogeneity at optimum)**.** $\frac{1}{M}\sum_{m=1}^M \|\nabla F_m(\mathbf{x}^\star)\|_2^2 \leq \zeta_*^2$.

Remark 3.2. *While the right measure of heterogeneity is a subject of significant debate in the FL community, the most popular are either a bound on gradient heterogeneity at $\mathbf{x}^\star$ (Assumption 3), or a stronger assumption of uniform gradient heterogeneity: for any $\mathbf{x}$, $\frac{1}{M}\sum_{m=1}^M \|\nabla F_m(\mathbf{x})\|_2^2 \leq \zeta^2$. The best-known lower bound, due to Woodworth et al. [2020a], considers the weaker Assumption 3. We remark however that most upper bounds use the stronger uniform assumption (e.g., [Khaled et al., 2020]).*

Theorem 3.3 (Lower bound for heterogeneous FEDAVG (see Theorem D.1)). *For any $K \geq 2, R, M$, H, D, σ, and ζ_*, there exist $f(\mathbf{x}; \xi)$ and distributions $\{\mathcal{D}_m\}$, each satisfying Assumption 1, and together satisfying Assumption 3, such that for some initialization $\mathbf{x}^{(0,0)}$ with $\|\mathbf{x}^{(0,0)} - \mathbf{x}^\star\|_2 < D$, the final iterate of* FEDAVG *with any step size satisfies:*

$$\mathbb{E}\left[F(\mathbf{x}^{(R,0)})\right] - F(\mathbf{x}^\star) \geq \Omega\left(\frac{HD^2}{KR} + \frac{\sigma D}{\sqrt{MKR}} + \min\left\{\frac{\sigma D}{\sqrt{KR}}, \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right\} + \min\left\{\frac{\zeta_*^2}{H}, \frac{H^{\frac{1}{3}}\zeta_*^{\frac{2}{3}}D^{\frac{4}{3}}}{R^{\frac{2}{3}}}\right\}\right)$$

Theorem 3.3 is nearly tight, up to a difference in the definitions of heterogeneity (See Remark 3.2). We compare our result to existing lower bounds and upper bounds in Table 1.

4 Upper Bounds for FEDAVG Under Third-Order Smoothness

In light of the limitations of FEDAVG discussed in Section 3, it is natural to ask if there are additional assumptions under which FEDAVG may perform better. Several classes of additional assumptions have been suggested for studying the performance of FEDAVG. Perhaps the most common, and the one supported from our intuition on the bias, is an assumption of third-order smoothness, stated formally in Assumption 2. Previously it has been shown that under such an assumption, FEDAVG may converge faster. We present several state-of-the-art bounds for FEDAVG under Assumption 2, including for the non-convex case.

Theorem 4.1 (Upper bound for FEDAVG under 3rd order smoothness (see Theorem E.1)). *Suppose $f(\mathbf{x}, \xi)$ satisfies Assumptions 1 and Assumptions 2. Then for some step size,* FEDAVG *satisfies*

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}})\|^2\right] \leq \mathrm{O}\left(\frac{HB}{KR} + \frac{\sigma\sqrt{BH}}{\sqrt{MKR}} + \frac{B^{\frac{4}{5}}\sigma^{\frac{4}{5}}Q^{\frac{2}{5}}}{K^{\frac{2}{5}}R^{\frac{4}{5}}}\right)$$

where $\hat{\mathbf{x}} := \frac{1}{M}\sum_m \mathbf{x}_m^{(r,k)}$ for a random choice of $k \in [K]$, and $r \in [R]$, and $B := F(\mathbf{x}^{(0,0)}) - \inf_{\mathbf{x}} F(\mathbf{x})$.

In the non-convex setting, as is standard in the FL literature [Stich, 2019, Yu et al., 2019b, Reddi et al., 2021], we require an assumption bounding moments of the stochastic gradients. Note that this is stronger that Assumption 1 which bounds the *variance* of the stochastic gradients.

Assumption 4 (Bounded gradients). *For any $\mathbf{x}$, we have $\mathbb{E}_\xi\left[\|\nabla f(\mathbf{x}, \xi)\|^4\right] \leq G^4$.*

Theorem 4.2 (Upper bound for FEDAVG with non-Convex objectives under third-order smoothness, see Theorem E.2). *Suppose $F(\mathbf{x})$ is H-smooth and $f(\mathbf{x}, \xi)$ satisfies Assumptions 2 and 4. Then for some step size, we have*

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}})\|^2\right] \leq \mathrm{O}\left(\frac{HB}{KR} + \frac{G\sqrt{BH}}{\sqrt{MKR}} + \frac{B^{\frac{4}{5}}G^{\frac{4}{5}}Q^{\frac{2}{5}}}{R^{\frac{4}{5}}}\right),$$

where $\hat{\mathbf{x}} := \frac{1}{M}\sum_m \mathbf{x}_m^{(r,k)}$ for a random choice of $k \in [K]$, and $r \in [R]$, and $B := F(\mathbf{x}^{(0,0)}) - \inf_{\mathbf{x}} F(\mathbf{x})$.

This theorem shows that the convergence rate of FEDAVG improves substantially under third order smoothness. In comparison, the best known rate for FEDAVG with non-convex objectives (under second-order smoothness alone) is $\frac{HB}{KR} + \frac{G\sqrt{BH}}{\sqrt{MKR}} + \frac{B^{\frac{2}{3}}G^{\frac{2}{3}}H^{\frac{2}{3}}}{R^{\frac{2}{3}}}$,due to Yu et al. [2019b]. Observe that we improve the dependence from $R^{\frac{2}{3}}$ in the third term to $R^{\frac{4}{5}}$.

5 Conclusion

In this work we provided sharp lower bounds for homogeneous and heterogeneous FEDAVG that matches the existing upper bound. By solving this open problem, we highlight the obstacles to FEDAVG, and show how a third-order smoothness assumption can lead to faster convergence. We expect the proposed techniques can shed light on the analysis of other federated algorithms and aid design of more efficient federated algorithms.

Acknowledgements

We would like to thank Aaron Sidford for helpful discussions. MG acknowledges the support of NSF award DGE-1656518. HY is partially supported by the TOTAL Innovation Scholars program. TM acknowledges the support of Google Faculty Award, NSF IIS 2045685, the Sloan Fellowship, and JD.com. We would like to thank the anonymous reviewers for their suggestions and comments.

References

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.

Alekh Agarwal, John Langford, and Chen-Yu Wei. Federated Residual Learning. *arXiv:2003.12880 [cs, stat]*, 2020.

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*, 2021.

Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed Distillation for On-Device Learning. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

Zachary Charles and Jakub Konečný. On the Outsized Importance of Learning Rates in Local Update Methods. *arXiv:2007.00878 [cs, math, stat]*, 2020.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated Meta-Learning with Fast Convergence and Efficient Communication. *arXiv:1802.07876 [cs]*, 2019.

Hong-You Chen and Wei-Lun Chao. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In *International Conference on Learning Representations*, 2021.

Xiangyi Chen, Tiancong Chen, Haoran Sun, Steven Z. Wu, and Mingyi Hong. Distributed Training with Heterogeneous Data: Bridging Median- and Mean-Based Algorithms. In *Advances in Neural Information Processing Systems 33*, 2020.

Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive Personalized Federated Learning. *arXiv:2003.13461 [cs, stat]*, 2020.

Enmao Diao, Jie Ding, and Vahid Tarokh. Hetero{FL}: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for Local-SGD with large step size. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *Annals of Statistics*, 48(3), 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman E. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems 33*, 2020.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Filip Hanzely and Peter Richtárik. Federated Learning of a Mixture of Global and Local Models. *arXiv:2002.05516 [cs, math, stat]*, 2020.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. In *Advances in Neural Information Processing Systems 33*, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. *arXiv:1909.06335 [cs, stat]*, 2019.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the 31st Conference on Learning Theory*, volume 75. PMLR, 2018.

Stanisław Jastrzkebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving Federated Learning Personalization via Model Agnostic Meta Learning. *arXiv:1909.12488 [cs, stat]*, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the International Conference on Machine Learning 1 Pre-Proceedings (ICML 2020)*, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter Theory for Local SGD on Identical and Heterogeneous Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108. PMLR, 2020.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, 1992.

Jakub Konecny, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. In *8th NIPS Workshop on Optimization for Machine Learning*, 2015.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3), 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020*, 2020b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. In *International Conference on Learning Representations*, 2020c.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance Reduced Local SGD with Lower Communication Complexity. *arXiv:1912.12844 [cs, math, stat]*, 2019.

Sen Lin, Guang Yang, and Junshan Zhang. Real-Time Edge Intelligence in the Making: A Collaborative Learning Framework via Federated Meta-Learning. *arXiv:2001.03229 [cs, stat]*, 2020a.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Advances in Neural Information Processing Systems 33*, 2020b.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations*, 2020c.

Ben London. PAC Identifiability in Federated Personalization. In *NeurIPS 2020 Workshop on Scalability, Privacy and Security in Federated Learning (SpicyFL)*, 2020.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., 2009.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54. PMLR, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Bernt Øksendal. *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003.

Reese Pathak and Martin J. Wainwright. FedSplit: An algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive Federated Optimization. In *International Conference on Learning Representations*, 2021.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Jonathan D. Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference*, 5(4), 2016.

Ohad Shamir and Nathan Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2014.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Canh T. Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized Federated Learning with Moreau Envelopes. In *Advances in Neural Information Processing Systems 33*, 2020.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv:1808.07576 [cs, stat]*, 2018.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Pengfei Wang, Risheng Liu, Nenggan Zheng, and Zhefeng Gong. Asynchronous Proximal Stochastic Gradient Algorithm for Composition Optimization Problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.

Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs Local SGD for Heterogeneous Distributed Learning. In *Advances in Neural Information Processing Systems 33*, 2020a.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020b.

Blake Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. *arXiv preprint arXiv:2102.01583*, 2021.

Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. FedMix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2021.

Hao Yu and Rong Jin. On the computation and communication complexity of parallel SGD with dynamic batch sizes for stochastic non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019a.

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, 2019b.

Honglin Yuan and Tengyu Ma. Federated Accelerated Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 33*, 2020.

Honglin Yuan, Warren Morningstar, Lin Ning, and Karan Singhal. What Do We Mean by Generalization in Federated Learning? *arXiv:2110.14216 [cs, stat]*, 2021a.

Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated Composite Optimization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021b.

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:2005.11418 [cs, stat]*, 2020.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.