z-SignFedAvg: A Unified Stochastic Sign-Based Compression for Federated Learning

Zhiwei Tang[1,2], **Yanmeng Wang**[1,2], **Tsung-Hui Chang**[1,2]

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[2]Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract

Federated Learning (FL) is a promising privacy-preserving distributed learning paradigm but suffers from high communication cost when training large-scale machine learning models. Sign-based methods, such as SignSGD, have been proposed as a biased gradient compression technique for reducing the communication cost. However, sign-based algorithms could diverge under heterogeneous data, which thus motivated the development of advanced techniques, such as the error-feedback method and stochastic sign-based compression, to fix this issue. Nevertheless, these methods still suffer from slower convergence rates, and none of them allows multiple local SGD updates like FedAvg. In this paper, we propose a novel noisy perturbation scheme with a general symmetric noise distribution for sign-based compression, which not only allows one to flexibly control the bias-variance tradeoff for the compressed gradient, but also provides a unified viewpoint to existing stochastic sign-based methods. More importantly, the proposed scheme enables the development of the very first sign-based FedAvg algorithm (z-SignFedAvg) to accelerate the convergence. Theoretically, we show that z-SignFedAvg achieves a faster convergence rate than existing sign-based methods and, under the uniformly distributed noise, can enjoy the same convergence rate as its uncompressed counterpart. Extensive experiments are conducted to demonstrate that the z-SignFedAvg can achieve competitive empirical performance on real datasets and outperforms existing schemes.

Introduction

We consider the Federated Learning (FL) network with one parameter server and n clients (McMahan et al. 2017; Li et al. 2020), with the focus on solving the following distributed learning problem

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

where $f_i(\cdot)$ is the local objective function for the i-th client, for $i = 1, \ldots, n$. Throughout this paper, we assume that each f_i is smooth and possibly non-convex. The local objective functions are generated from the local dataset owned by each client. When designing distributed algorithms to solve (1), a crucial aspect is the communication efficiency since a massive number of clients need to transmit their local gradients to the server frequently (Li et al. 2020). As one of the most popular FL algorithms, the federated averaging (FedAvg) algorithm (McMahan et al. 2017; Konečnỳ et al. 2016) considers multiple local SGD updates with periodic communications to reduce the communication cost. Another way is to compress the local gradients before sending them to the server (Li et al. 2020; Alistarh et al. 2017; Reisizadeh et al. 2020). Among the existing compression methods, a simple yet elegant technique is to take the sign of each coordinate of the local gradients, which requires only one bit for transmitting each coordinate. For any $x \in \mathbb{R}$, we define the sign operator as: $\text{Sign}(x) = 1$ if $x \geq 0$ and -1 otherwise.

It has been shown recently that optimization algorithms with the sign-based compression can enjoy a great communication efficiency while still achieving comparable empirical performance as uncompressed algorithms (Bernstein et al. 2018; Karimireddy et al. 2019; Safaryan and Richtárik 2021). However, for distributed learning, especially the scenarios with heterogeneous data, i.e., $f_i \neq f_j$ for every $i \neq j$, a naive application of the sign-based algorithm may end up with divergence (Karimireddy et al. 2019; Chen et al. 2020; Safaryan and Richtárik 2021).

A counterexample for sign-based distributed gradient descent. Consider the one-dimensional problem with two clients: $\min_{x\in\mathbb{R}} (x-A)^2 + (x+A)^2$, where $A > 0$ is some constant. For any $x \in [-A, A]$, the averaged sign gradient at x is $\text{Sign}(x-A) + \text{Sign}(x+A) = 0$, i.e., the algorithm never moves. Similar examples are also discussed by (Chen et al. 2020; Safaryan and Richtárik 2021). The fundamental reason for this undesirable result is the uncontrollable bias brought by the sign-based compression.

There are mainly two approaches to fixing this issue in the existing literature. The first one is the stochastic sign-based method, which introduces stochasticity into the sign operation (Jin et al. 2020; Safaryan and Richtárik 2021; Chen et al. 2020), and the second one is the Error-Feedback (EF) method (Karimireddy et al. 2019; Vogels, Karimireddy, and Jaggi 2019; Tang et al. 2019). However, these works are still unsatisfactory. Specifically, on one hand, both the theoretical convergence rates and empirical performance of these algorithms are still worse than uncompressed algorithms like (Ghadimi and Lan 2013; Yu, Yang, and Zhu 2019). On the

other hand, none of them allows the clients to have multiple local SGD updates within one communication round like the FedAvg, which thereby are less communication efficient. This work aims at addressing these issues and closing the gaps for sign-based methods.

Main contributions. Our contributions are summarized as follows.

(1) **A unified family of stochastic sign operators.** We show an intriguing fact: The bias brought by the sign-based compression can be flexibly controlled by injecting a proper amount of random noise before the sign operation. In particular, our analysis is based on a novel noisy perturbation scheme with a general symmetric noise distribution, which also provides a unified framework to understand existing stochastic sign-based methods including (Jin et al. 2020; Safaryan and Richtárik 2021; Chen et al. 2020).

(2) **The first sign-based FedAvg algorithm.** In contrast to the existing sign-based methods which do not allow multiple local SGD updates within one communication round, based on the proposed stochastic sign-based compression, we design a novel family of sign-based federated averaging algorithms (z-SignFedAvg) that can achieve the best of both worlds: high communication efficiency and fast convergence rate.

(3) **New theoretical convergence rate analyses.** By leveraging the asymptotic unbiasedness property of the stochastic sign-based compression, we derive a series of theoretical results for z-SignFedAvg and demonstrate its improved convergence rates over the existing sign-based methods. In particular, we show that by injecting a sufficiently large uniform noise, z-SignFedAvg can have a matching convergence rate with the uncompressed algorithms.

Notations. For any $x \in \mathbb{R}^d$, we denote $x(j)$ as the j-th element of the vector x. We define the ℓ_p-norm for $p \geq 1$ as $\|x\|_p = (\sum_{j=1}^d |x(j)|^p)^{\frac{1}{p}}$. We denote that $\|\cdot\| = \|\cdot\|_2$, and $\|x\|_\infty = \max_{j\in\{1,...,d\}} |x(j)|$. For any function $f(x)$, we denote $f^{(k)}(x)$ as its k-th derivative, and for a vector $x = [x(1), ..., x(d)]^\top \in \mathbb{R}^d$, we define $\text{Sign}(x) = [\text{Sign}(x(1)), ..., \text{Sign}(x(d))]^\top$.

Related Works

Stochastic sign-based method. Our proposed algorithm belongs to this category. Among the existing works (Safaryan and Richtárik 2021; Jin et al. 2020; Chen et al. 2020), the setting considered by (Safaryan and Richtárik 2021) is closest to ours since the latter two consider gradient compression not only in the uplink but also in the downlink. Despite of this difference and the use of different convergence metrics, the algorithms therein achieve the same convergence rate $O(\tau^{-\frac{1}{4}})$, where τ is the total number of gradient queries to the local objective function. Compared to existing works, our proposed z-SignFedAvg requires a slightly stronger assumption on the minibatch gradient noise, but achieves a faster convergence rate $O(\tau^{-\frac{1}{3}})$ or even $O(\tau^{-\frac{1}{2}})$, with the standard squared ℓ_2-norm of gradients as the convergence metric.

Error-Feedback method. The error-feedback (EF) method is first proposed by (Seide et al. 2014) and later theoretically justified by (Karimireddy et al. 2019). Then, (Vogels, Karimireddy, and Jaggi 2019; Tang et al. 2019, 2021a) further extended this EF method into distributed and adaptive gradient schemes. The key idea of the EF-based methods is to show that the sign operator scaled by the gradient norm is a contractive compressor, and the error induced by the contractive compressor can be compensated. However, such EF-based methods cannot deal with partial client participation otherwise the error residuals cannot be correctly tracked. Besides, the EF-based methods have a convergence rate $\mathcal{O}(\tau^{-\frac{1}{2}} + d^2\tau^{-1})$, where d is the dimension of the gradients, and therefore is not competitive for high-dimension problems.

Unbiased quantization method. Apart from the sign-based gradient compression, another popular way of compression is the unbiased stochastic quantization method adopted by (Alistarh et al. 2017; Reisizadeh et al. 2020; Haddadpour et al. 2021; Vargaftik et al. 2021). A key assumption made by this category of methods is that the quantization error is bounded by the norm of the input, which however does not hold for sign-based compression, and therefore the existing convergence results therein do not apply to sign-based methods. Besides, as shown in (Alistarh et al. 2017; Reisizadeh et al. 2020; Vargaftik et al. 2021), these unbiased methods usually have degraded convergence speed when compared to the uncompressed algorithms.

As mentioned, some of the existing sign-based methods like (Chen et al. 2020; Safaryan and Richtárik 2021) do not adopt the standard squared ℓ_2-norm of gradients as the metric for the convergence rate analysis. Thus, it is tricky to make a fair comparison between them and the proposed z-SignFedAvg. We provide a detailed discussion in the Appendix to summarize the convergence rates of some representative algorithms.

Sign Operator with Symmetric and Zero-Mean Noise

In this section, we introduce a general noisy perturbation scheme for the sign-based compression and analyze the asymptotic unbiasedness of compressed gradients. The results serve as the foundation for the proposed algorithms in subsequent sections.

Key observation. Let ξ be a random variable that is symmetric, zero-mean and has the p.d.f $p(t)$. If $p(0) \neq 0$ and $p(t)$ is continuous and uniformly bounded on $(-\infty, +\infty)$, then it can be verified that

$$\lim_{\sigma\to+\infty} \frac{\sigma}{2p(0)} \mathbb{E}[\text{Sign}(x + \sigma\xi)] = x. \tag{2}$$

In other words, the perturbed sign operator is an asymptotically unbiased estimator of the input x when $\sigma \to \infty$. Furthermore, assume that $p(t)$ is uniformly bounded on $(-\infty, +\infty)$ and differentiable for an arbitrary order. Then, with the Tay-

lor's expansion, we can have

$$\frac{\sigma}{p(0)}\int_0^{\frac{x}{\sigma}} p(t)dt = x + \frac{1}{p(0)}\sum_{k=1}^{+\infty}\frac{p^{(k)}(0)x^{k+1}}{(k+1)!\sigma^k}$$
$$= x + \sum_{k=1}^{+\infty} p^{(k)}(0)\mathcal{O}\left(\sigma^{-k}\right).$$

Therefore, suppose that K is the largest integer such that $p^{(1)}(0) = 0, ..., p^{(K)}(0) = 0$. The LHS of (2) will converge to x with the order $\mathcal{O}(\sigma^{-(K+1)})$. This observation motivates us to find a distribution with $p^{(i)}(0) = 0$ for all $i \leq z$, given a positive integer $z \in \mathbb{Z}_+$, which leads to the following family of noise distribution parameterized by z.

Definition 1 (z-distribution)**.** *A random variable ξ_z is said to follow the z-distribution if its p.d.f is*

$$p_z(t) = \frac{1}{2\eta_z}e^{-\frac{t^{2z}}{2}}, \tag{3}$$

where $\eta_z = 2^{\frac{1}{2z}}\Gamma\left(1+\frac{1}{2z}\right)$ and $\Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}dt$ is the Gamma function.

It can be verified that $p_z(t)$ in (3) is a valid p.d.f. When $z = 1$, it corresponds to the standard Gaussian distribution. In addition, one can also show that $p_z(t)$ converges to the p.d.f of the uniform distribution when $z \to +\infty$, as summarized in Lemma 1 below.

Lemma 1. *The z-distribution weakly converges to uniform distribution on $[-1, 1]$ when $z \to +\infty$.*

This z-distribution has a nice property that can be leveraged to bound the bias caused by the sign-based compression, as stated in the following lemma.

Lemma 2. *For any $x \in \mathbb{R}^d$ and $\sigma > 0$,*

$$\left\|\eta_z\sigma\mathbb{E}\left[\mathrm{Sign}(\mathrm{x}+\sigma\xi_z)\right] - x\right\|^2 \leq \frac{\|x\|_{4z+2}^{4z+2}}{4(2z+1)^2\sigma^{4z}}, \tag{4}$$

where $\xi_z(1), ..., \xi_z(d)$ follow the i.i.d. z-distribution.

Remark 1. *One can see that the RHS of* (4) *involves the term $(\|x\|_{4z+2}/\sigma)^{4z}$. Thus, as long as $\sigma > \|x\|_\infty$, the LHS of* (4) *converges to zero when $z \to +\infty$. Since Lemma 1 implies that ξ_∞ follows the i.i.d uniform distribution on $[-1, 1]$, we obtain $\sigma\mathbb{E}\left[\mathrm{Sign}(\mathrm{x}+\sigma\xi_\infty)\right] = x$ as long as $\sigma > \|x\|_\infty$. It is interesting to remark that the stochastic sign operators proposed in (Jin et al. 2020; Safaryan and Richtárik 2021) are exactly the sign operator injected by the uniform noise, and (Chen et al. 2020) also considered the use of a symmetric noise for gradient perturbation. Thus, sign-based compression with the z-distribution offers a unified perspective to understand the relationship among the existing stochastic sign-based methods.*

z-SignFedAvg Algorithm

In this section, we propose the following sign-based FedAvg algorithm, termed as z-SignFedAvg. While FedAvg-type algorithms with gradient compression are also presented in (Haddadpour et al. 2021), they require unbiased compression and are not applicable to sign-based methods. The details of z-SignFedAvg are presented in Algorithm 1. A prominent difference between the proposed z-SignFedAvg and the existing sign-based methods lies in that the clients are allowed to perform multiple SGD updates per communication round ($E > 1$) before applying the stochastic sign-based compression. Like the FedAvg algorithm, it is anticipated that z-SignFedAvg can greatly benefit from this and has a significantly reduced communication cost.

Note that in practice we only consider $z = 1$ and $z = +\infty$ for the z-SignFedAvg since they correspond to the Gaussian distribution and uniform distribution, respectively. Nevertheless, we are interested in the convergence properties of z-SignFedAvg for a general positive integer z as it provides better insights on the role of z for the convergence rate.

Algorithm 1: z-SignFedAvg (or z-SignSGD when $E = 1$)

Require: Total communication rounds T, number of local steps E, number of clients n, clients stepsize γ, server stepsize η, noise coefficient σ, parameter of noise distribution z.

Initialize x_0.
for $t = 1$ to T **do**
  On Clients:
  for $i = 1$ to n **do**
    $x^i_{t-1,0} = x_{t-1}$
    for $s = 1$ to E **do**
      $g^i_{t-1,s} = g_i(x^i_{t-1,s-1})$, where $g_i(\cdot)$ is the minibatch gradient oracle of the i-th client.
      $x^i_{t-1,s} = x^i_{t-1,s-1} - \gamma g^i_{t-1,s}$.
    end for
    Sample $\xi_z \in \mathbb{R}^d$ from the distribution $p_z(t)$ i.i.d.
    $\Delta^i_{t-1} = \mathrm{Sign}\left(\frac{x_{t-1}-x^i_{t-1,E}}{\gamma} + \sigma\xi_z\right)$.
    Send Δ^i_{t-1} to the server.
  end for
  On Server:
  $x_t = x_{t-1} - \eta\gamma\frac{1}{n}\sum_{i=1}^n \Delta^i_{t-1}$.
  Broadcast x_t to the clients.
end for

We first state some standard assumptions for problem (1).

Assumption 1. *We assume that each $f_i(x)$ has the following properties:*

A.1 The minibatch gradient is unbiased and has bounded variance, i.e., $\mathbb{E}[g_i(x)] = \nabla f_i(x)$ and

$$\mathbb{E}[\|g_i(x) - \nabla f_i(x)\|_2^2] \leq \zeta^2.$$

A.2 Each f_i is smooth, i.e., for any $x, y \in \mathbb{R}^d$, there exists some non-negative constants $L_1, \ldots, L_d$, such that

$$f(y) - f(x) \leq \langle\nabla f(x), y - x\rangle + \frac{\sum_{j=1}^d L_j\left(y(j) - x(j)\right)^2}{2}.$$

A.3 f is lower bounded, i.e., there exists some constant f^* such that $f(x) \geq f^*, \forall x \in \mathbb{R}^d$.

A.4 There exists a constant $G \geq 0$ such that $\|\nabla f_i(x)\| \leq G$, $\forall i = 1, ..., n$, and $x \in \mathbb{R}^d$.

Assumption A.2 is a more fine-grained assumption on the function smoothness than the commonly used one and is also

used by (Bernstein et al. 2018; Safaryan and Richtárik 2021). For the convergence rate analysis, we consider two cases, namely, the case with $z < +\infty$ and the case of $z = \infty$.

Case 1: $z < +\infty$

As we can see from Lemma 2, there always exists some gradient bias when $z < +\infty$. In order to bound it, we further assume that a higher order moment of the minibatch gradient noise is bounded.

Assumption 2. *There exists a constant $Q_z \geq 0$ such that for any $x \in \mathbb{R}^d$, we have*

$$\mathbb{E}[\|g_i(x) - \nabla f_i(x)\|_{4z+2}^{4z+2}] \leq Q_z. \tag{5}$$

Theorem 1. *Suppose that Assumption 1 and 2 hold. Denote $\bar{x}_{t,s} = \frac{1}{n}\sum_{i=1}^n x_{t,s}^i$ and $L_{\max} = \max_j L_j$. Then, for $\eta = \eta_z\sigma$, $\gamma \leq \frac{1}{L_{\max}}$ and $z < +\infty$ in Algorithm 1, we have*

$$\mathbb{E}\left[\frac{1}{TE}\sum_{t=1}^{T}\sum_{s=1}^{E}\|\nabla f(\bar{x}_{t-1,s-1})\|^2\right]$$

$$\leq \underbrace{\frac{2\mathbb{E}[f(x_0) - f^*]}{TE\gamma} + \frac{\gamma\zeta^2 L_{\max}}{n} + \frac{4\gamma^2(E-1)EL_{\max}^2(\zeta^2+G^2)}{3}}_{\text{(a) Standard terms in FedAvg}} \tag{6a}$$

$$+ \underbrace{\frac{2^{2z+1}E^{2z}\sqrt{Q_z + G^{4z+2}}G}{\sqrt{2}(2z+1)\sigma^{2z}} + \frac{\gamma 2^{4z}E^{4z+1}(Q_z+G^{4z+2})L_{\max}}{2(2z+1)^2\sigma^{4z}}}_{\text{(b) Bias terms}} \tag{6b}$$

$$+ \underbrace{\frac{4\eta_z^2\gamma\sigma^2\sum_{j=1}^d L_j}{En}}_{\text{(c) Variance term}}. \tag{6c}$$

When is the bound non-trivial? Since we assume that the ℓ_2-norm of gradient is bounded by G, all the terms in the RHS of (6) should be no larger than G^2. For example, to have the first term in (6b) less than G^2, one requires σ to be greater than $2^{1+\frac{1}{4z}}E\left(Q_z/G^2 + G^{4z}\right)^{\frac{1}{4z}}/(2z+1)^{\frac{1}{2z}}$.

Bias-variance trade-off. An interesting observation from Theorem 1 is that there exists a trade-off between the bias and variance terms. One can see that the terms in (6b) is caused by the gradient bias of the sign operation (see (4)) and is an infinitesimal of σ with $\mathcal{O}\left(\sigma^{-2z}\right)$, while the term in (6c) is due to the injected noise and is in the order of $\mathcal{O}\left(\gamma\sigma^2\right)$. Specifically, the first term in (6b) only depends on the noise scale σ and mostly affects the final objective. Meanwhile, the variance term in (6c) mainly affects the convergence speed because a smaller stepsize is required for it to diminish.

Theoretically, we can choose an iteration-dependent noise scale σ so as to make the algorithm converge to a stationary solution. To see this, let us denote $\tau = TE$ as the total number of gradient queries per client, and present the following corollary.

Corollary 1 (Informal)**.** *Let $\sigma = (n\tau)^{\frac{1}{4z+2}}$ and $\gamma = \min\{n^{\frac{z}{2z+1}}\tau^{-\frac{z+1}{2z+1}}, L_{\max}^{-1}\}$ in Theorem 1, and let $E \leq n^{-\frac{3z}{4z+2}}\tau^{\frac{z+2}{4z+2}}$. We have*

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{T}\sum_{s=1}^{E}\|\nabla f(\bar{x}_{t-1,s-1})\|^2\right] = \mathcal{O}\left((n\tau)^{-\frac{z}{2z+1}}\right). \tag{7}$$

Achieveing linear speedup. From Corollary 1, we can see that the z-SignFedAvg needs $(n\tau)^{\frac{3z}{4z+2}}$ communication rounds to achieve a linear-speedup convergence rate. Particularly, when $z = 1$, the corresponding convergence rate is $\mathcal{O}((n\tau)^{-\frac{1}{3}})$ and the required communication rounds is $(n\tau)^{\frac{1}{2}}$. To the best of our knowledge, the previous works have never shown the sign-based method can achieve a linear-speedup convergence rate.

Relationship to (Chen et al. 2020). The work (Chen et al. 2020) also considered the use of a symmetric and zero-mean noise for the sign-based compression and proved that the algorithm has a convergence rate $\mathcal{O}(\tau^{-\frac{1}{4}})$. However, their results have three differences from our z-SignFedAvg and Theorem 1. First, (Chen et al. 2020) considered gradient compression both in the uplink and downlink communications. In addition, the convergence metric they used is not the standard squared ℓ_2-norm of gradients and is hard to interpret. Second, their analysis is rooted in the median-based algorithm, whereas we judiciously exploit the property of the sign operation and hence provide a general analysis framework for the stochastic sign-based methods. Last but not the least, unlike our z-SignFedAvg, (Chen et al. 2020) cannot allow multiple local SGD updates.

Case 2: $z = +\infty$

When $z = +\infty$, the injected noise ξ_z in the z-SignFedAvg is uniformly distributed on $[-1, 1]$. From Remark 1, we have learned that the gradient bias can vanish as long as the noise scale σ is sufficiently large. To quantify this threshold, we need the following assumption which is a limit form of Assumption 2.

Assumption 3. *There exists a constant $Q_\infty \geq 0$ such that for any $x \in \mathbb{R}^d$, with probability 1,*

$$\|g_i(x) - \nabla f_i(x)\|_\infty \leq Q_\infty. \tag{8}$$

Theorem 2. *(Informal) Suppose that Assumption 1 and 3 hold. For $\gamma = \min\{n^{\frac{1}{2}}\tau^{-\frac{1}{2}}, L_{\max}^{-1}\}$, $\eta = \sigma$, $z = +\infty$, $E \leq n^{-\frac{3}{4}}\tau^{\frac{1}{4}}$ and $\sigma > E(G + Q_\infty)$ in Algorithm 1 we have*

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{T}\sum_{s=1}^{E}\|\nabla f(\bar{x}_{t-1,s-1})\|^2\right] = \mathcal{O}\left((n\tau)^{-\frac{1}{2}}\right). \tag{9}$$

However, if $\sigma \leq E(G+Q_\infty)$, there exists a problem instance for which Algorithm 1 cannot converge.

Remark 2. *Note that Theorem 2 implies that ∞-SignFedAvg has a matching convergence rate as the uncompressed FedAvg. The reason why ∞-SignFedAvg cannot converge when $\sigma \leq E(G + Q_\infty)$ is simply that the uniform noise has a finite support and cannot always change the sign of gradients. For example, if $\sigma < A$ for some $A > 0$, then we have $\mathrm{Sign}(x + \sigma\xi_\infty) = \mathrm{Sign}(x)$ for any $x \geq A$.*

Relationship to (Jin et al. 2020; Safaryan and Richtárik 2021). As mentioned in Remark 1, both the stochastic sign operators in (Jin et al. 2020; Safaryan and Richtárik 2021) are equivalent to the sign operator injected by the uniform noise. Nevertheless, there are still two distinctions when compared with our ∞-SignFedAvg. First, while (Safaryan and Richtárik 2021) shows their algorithm has a $\mathcal{O}(\tau^{-\frac{1}{4}})$ convergence rate, it is based on the ℓ_2-norm of gradients and cannot imply the same rate as that in (9) (see Appendix for more details). Second, although (Safaryan and Richtárik 2021) does not need Assumption 3, it relies on an input-dependent noise scale which, unfortunately, often slows the algorithm convergence in practice especially when the problem dimension is large.

Case	Rate	Threshold on σ	Assumption
1	$\mathcal{O}(\tau^{-\frac{z}{2z+1}})$	$\widetilde{\mathcal{O}}\left(\left(\frac{Q_z}{G^2}+G^{4z}\right)^{\frac{1}{4z}}\right)$	Asp. 2
2	$\mathcal{O}(\tau^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}(Q_\infty+G)$	Asp. 3

Table 1: Comparison of Case 1 and Case 2.

More theoretical results and proofs are relegated to Appendix. Below, we have two more remarks.

Remark 3. ***(Bounded minibatch gradient noise)*** *While both Assumption 2 and 3 are slightly stronger than the commonly used second-order condition on the minibatch gradient noise, they are still justifiable since unbounded minibatch gradient noise is rarely to happen in practice.*

Remark 4. ***(Minibatch gradient noise works as noise perturbation)*** *When the minibatch gradient is used as the input of the sign operator in* (2)*, the minibatch gradient noise itself may function as the perturbation noise. In particular, as shown in (Chen, Wu, and Hong 2020) the minibatch gradient noise approximately follows a symmetric distribution. Therefore, in practice, one may not need to inject as large noises as suggested by Theorem 2 since the minibatch gradient noise can also help mitigate the bias due to sign-based compression. This also explains why a small noise scale is sufficient for z-SignFedAvg to achieve good performance in the experiment section.*

Comparison of Case 1 and Case 2

We summarize the results of Case 1 and Case 2 in Table 1, where $\widetilde{O}(\cdot)$ hides some constants that do not affect the comparison. Especially, we can see that when the mini-batch gradient noise has a long tail such that $Q_z/G^2 \ll Q_\infty^{4z}$, Case 1 requires a less amount of noise than Case 2 for guaranteeing convergence. Despite of the difference in theory, we will see in the experiment section that z-SignFedAvg under Case 1 and Case 2 have almost the same behavior in practice.

Experiments

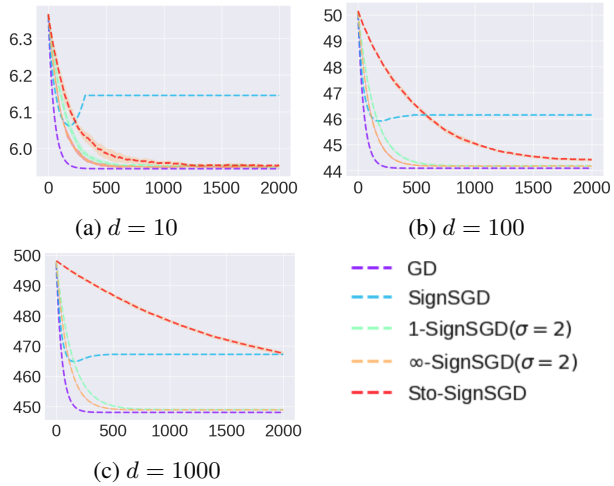
In this section, we present the experiment results on both synthetic and real problems, and all the figures in this section are obtained by 10 independent runs and are visualized in the form of mean$\pm$std.

Noise scale as a hyperparameter. Although we explicitly characterize how the performance of z-SignFedAvg depends on the noise scale σ in the previous section, we treat σ as a tunable hyperparameter in the experiments. This is because, on one hand, the theoretical lower bound for σ are difficult to compute since it is impossible to access the moment condition of the minibatch gradient noise. On the other hand, as we have discussed in Remark 4, owing to the presence of the minibatch gradient noise, we can use a much smaller noise scale than the theoretical one in practice.

A Simple Consensus Problem

In this section, we verify our previous theoretical results by considering the simple consensus problem with 10 clients: $\min_{x\in\mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{10}\|x-y_i\|^2$, where $y_1,...,y_{10}\in\mathbb{R}^d$ are generated using i.i.d standard Gaussian distribution, and d is the problem dimension. We implemented the following algorithms: GD (Gradient descent), Sto-SignSGD (Safaryan and Richtárik 2021), SignSGD (Algorithm 1 with $z=1$, $E=1$ and $\sigma=0$), 1-SignSGD (Algorithm 1 with $z=1$ and $E=1$.), ∞-SignSGD (Algorithm 1 with $z=+\infty$ and $E=1$). For all the algorithms, we considered the full gradient (no minibatch SGD), and used the same stepsize 0.01 and initialization by a zero vector.

Results. As we can see from Figure 1, the vanilla SignSGD fails to converge to the optimal solution whereas the others can. Besides, 1-SignSGD and ∞-SignSGD have roughly the same convergence speed which is slightly slower than the uncompressed GD. It is also observed that the input-dependent noise scale adopted by (Safaryan and Richtárik 2021) could slow the convergence when the problem dimension is high.

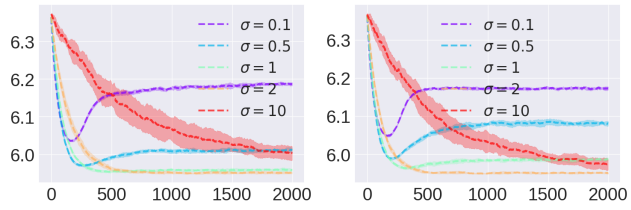


(a) $d=10$ (b) $d=100$

(c) $d=1000$

Figure 1: Performance of tested algorithms under different problem dimensions. The x-axis is communication rounds and the y-axis is the objective values.

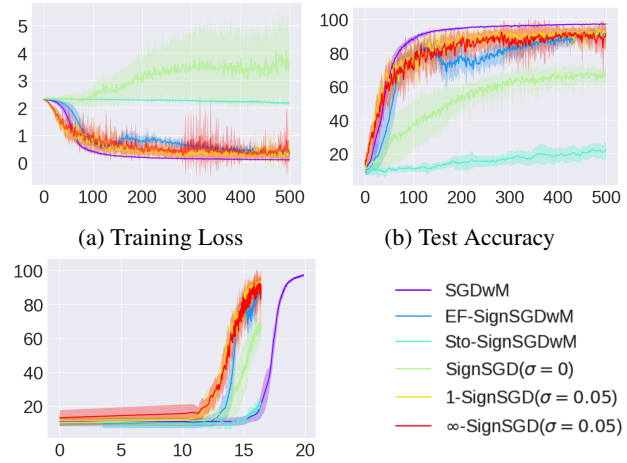
Figure 2 displays the results of 1-SignSGD and ∞-SignSGD with various noise scales. We can see that there is a clear bias-variance trade-off for different noise scales and



(a) 1-SignSGD (b) ∞-SignSGD

Figure 2: z-SignSGD under various noise scales.

it corroborates our analysis after Theorem 1. It is also worth mentioning that the best choice of σ for Algorithm 1 shown in Figure 2 is much smaller than the one predicted by the theorems.



(a) Training Loss (b) Test Accuracy

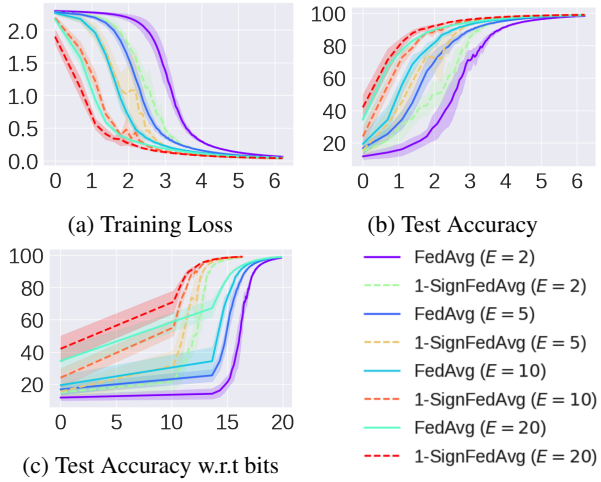
(c) Test Accuracy w.r.t bits

Figure 3: Performance of various SignSGD algorithms on non-i.i.d MNIST. x-axis: Communication rounds for (a) (b), Logarithmic of bits for (c).

z-SignSGD on Non-i.i.d MNIST

In this section, we consider an extremely non-i.i.d setting with the MNIST dataset (Deng 2012). Specifically, we split the dataset into 10 parts based on the labels and each client has the data of one digit only. A simple two-layer convolutional neural network (CNN) from Pytorch tutorial (Paszke et al. 2017) was used. The following algorithms were implemented: SGDwM (Distributed SGD (Ghadimi and Lan 2013) with momentum), EF-SignSGDwM (Distributed SignSGD with error-feedback and momentum (Karimireddy et al. 2019; Vogels, Karimireddy, and Jaggi 2019)), and Sto-SignSGDwM (Sto-SignSGD with momentum (Safaryan and Richtárik 2021)). For each of the algorithms, we selected its best hyperparameters, including the stepsize, momentum coefficient and the noise scale, via grid search (see Appendix).

Results. One can observe from Figure 3a-3b that again the vanilla SignSGD does not converge well. The proposed 1-SignSGD and ∞-SignSGD clearly outperform the existing EF-SignSGDwM and Sto-SignSGDw, and perform closely to the uncompressed SGDwM. The reason for the slow convergence of Sto-SignSGDw is that the injected noise is too large due to the input-dependent noise scale. Figure 3c further displays the testing accuracy of all methods versus the accumulated number of bits transmitted from the clients to the server. One can see that the proposed algorithms achieve the state-of-the-art performance on this task. More results for 1-SignSGD and ∞-SignSGD under different noise scales are presented in Appendix.



(a) Training Loss (b) Test Accuracy

(c) Test Accuracy w.r.t bits

Figure 4: Performance of FedAvg and 1-SignFedAvg on the EMNIST dataset. x-axis: Logarithmic of communication rounds for (a) (b), Logarithmic of bits for (c).

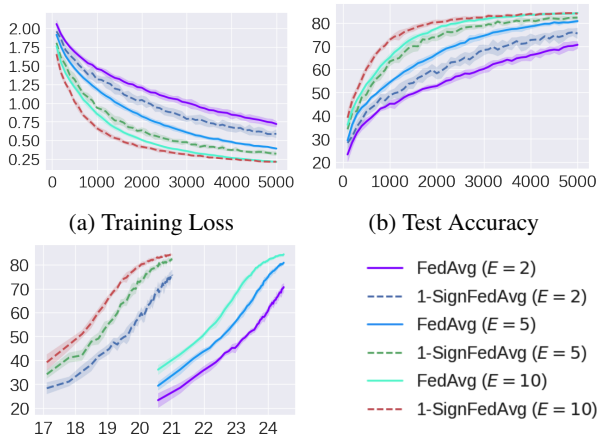
z-SignFedAvg on EMNIST and CIFAR-10

In this section, we evaluate the performance of our proposed z-SignFedAvg on two classical datasets: EMNIST(Cohen et al. 2017) and CIFAR-10 (Krizhevsky and Hinton 2010). In particular, the proposed z-SignFedAvg with $z = 1$ and $z = \infty$ are benchmarked against the uncompressed FedAvg (McMahan et al. 2017; Yu, Yang, and Zhu 2019). Since 1-SignFedAvg and ∞-SignFedAvg behave similarly, we only report the results of 1-SignFedAvg in this section and relegate the others to Appendix. For EMNIST, we use the same 2-layer CNN as the one in last experiment. For CIFAR-10, we used the ResNet18 (He et al. 2016) with group normalization (Wu and He 2018).

Settings. For both the experiments on EMNIST and CIFAR-10, we followed a setting similar to (Reddi et al. 2020). We also considered the scenario with partial client participation. For the EMNIST dataset, there are 3579 clients in total and 100 clients were uniformly sampled in each communication round to upload their compressed gradients. For the CIFAR-10 dataset, the training samples are partitioned among 100 clients, and each client has an associated multinomial distribution over labels drawn from a symmetric Dirichlet distribution with parameter 1. In each communication round, 10 out of 100 clients were uniformly sampled. We fixed the client stepsize as 0.05 and 0.1 for EMNIST

dataset and CIFAR-10 dataset respectively. For both dataset, we set the local batchsize as 32. The same noise scales for 1-SignFedAvg and ∞-SignFedAvg were used: $\sigma = 0.01$ for EMNIST and $\sigma = 0.0005$ for CIFAR-10. More details about the hyperparameters are referred to Appendix.

Results. We can see from Figure 5 that both uncompressed FedAvg and 1-SignFedAvg can benefit from multiple local SGD steps. More surprisingly, 1-SignFedAvg can even outperform the uncompressed FedAvg. This is probably because the EMNIST dataset is less heterogeneous than the one we used in the non-i.i.d MNIST. The results on the performance of 1-SignFedAvg and ∞-SignFedAvg under various choices of noise scales are relegated to Appendix, which are also consistent with our theoretical claims.



(a) Training Loss (b) Test Accuracy

(c) Test Accuracy w.r.t bits

Figure 5: Performance of FedAvg and 1-SignFedAvg on the CIFAR-10 dataset. x axis is the same as Figure 3.

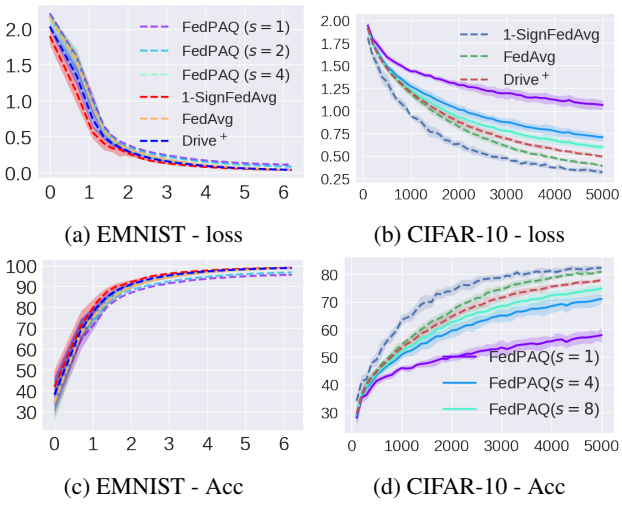
Comparison with Unbiased Stochastic Quantization Method

Aside from the experiments presented in this section, we also compare our algorithm to another popular family of unbiased stochastic compressed FL algorithms, namely, the FedPAQ in (Reisizadeh et al. 2020) and also the Drive$^+$ in (Vargaftik et al. 2021).

As we have shown that z-SignFedAvg with the Gaussian noise and uniform noise behave very closely, here we only consider 1-SignFedAvg for comparison.

Setting. Again, we consider the two FL datasets used in previous experiments. Specifically, we compare 1-SignFedAvg with FedPAQ on EMNIST and CIFAR-10. For all the algorithms, the client's stepsize and batchsize are set to the same values used in previous experiment. For the number of local steps, we set $E = 20$ for EMNIST and $E = 5$ for CIFAR-10. For 1-SignFedAvg, we reuse the previously found optimal hyperparameters. For FedPAQ and Drive$^+$, we tune the server stepsize via grid search on $[1, 0.5, 0.1, 0.05, 0.01, 0.005]$. The chosen hyperparameter FedPAQ and Drive$^+$ under three datesets are presented in Appendix.

Results. First, our result in Figure 6 is consistent to the result reported in (Vargaftik et al. 2021) on the EMNIST dataset, where Drive$^+$ can slightly outperform the uncompressed algorithm FedAvg. Secondly, our algorithm is superior to Drive$^+$ and the FedPAQ with low precision region (1 bit to 8 bits) on all datasets, more importantly, it dominates all the algorithms by a large margin particularly on the CIFAR-10 dataset. These results again, as (Bernstein et al. 2018; Karimireddy et al. 2019) did, show that the biased compressor, or more specifically the sign-based compressor, can be a strong competitor to those unbiased quantizer due to reduced variance.



(a) EMNIST - loss (b) CIFAR-10 - loss

(c) EMNIST - Acc (d) CIFAR-10 - Acc

Figure 6: Performance of FedAvg, 1-SignFedAvg, FedPAQ, Drive$^+$ on EMNIST/CIFAR-10.

Conclusion

In this work, we have proposed the z-SignFedAvg: a FedAvg-type algorithm with the stochastic sign-based compression. Thanks to the novel noisy perturbation scheme in Lemma 2, the proposed z-SignFedAvg provides a unified viewpoint to the existing sign-based methods as well as a general framework for convergence rate analysis. Through both theoretical analyses and empirical experiments, we have shown that the z-SignFedAvg can perform nearly the same, sometimes even better, than the uncompressed FedAvg and enjoy a significant reduction in the number of bits transmitted from clients to the server. As a final remark, the stochastic sign-based compression proposed in this work can be of independent interest and can be conveniently combined with other adaptive FL algorithms or gradient sparsification techniques such as those in (Karimireddy et al. 2020; Reddi et al. 2020; Basu et al. 2019), to further improve the communication efficiency.

Acknowledgments

The work is supported by Shenzhen Science and Technology Program under Grant No. RCJC20210609104448114, the NSFC, China, under Grant 62071409, and by Guangdong Provincial Key Laboratory of Big Data Computing.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016a. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283.

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016b. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Agarwal, N.; Kairouz, P.; and Liu, Z. 2021. The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34: 5052–5064.

Agarwal, N.; Suresh, A. T.; Yu, F. X. X.; Kumar, S.; and McMahan, B. 2018. cpSGD: Communication-efficient and differentially-private distributed SGD. *Advances in Neural Information Processing Systems*, 31.

Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; and Vojnovic, M. 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in neural information processing systems*, 30.

Amiri, S.; Belloum, A.; Klous, S.; and Gommans, L. 2021. Compressive Differentially-Private Federated Learning Through Universal Vector Quantization.

Asoodeh, S.; Liao, J.; Calmon, F. P.; Kosut, O.; and Sankar, L. 2021. Three variants of differential privacy: Lossless conversion and applications. *IEEE Journal on Selected Areas in Information Theory*, 2(1): 208–222.

Basu, D.; Data, D.; Karakus, C.; and Diggavi, S. 2019. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32.

Bernstein, J.; Wang, Y.-X.; Azizzadenesheli, K.; and Anandkumar, A. 2018. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 560–569. PMLR.

Bertsekas, D. P. 1997. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334.

Bu, Z.; Dong, J.; Long, Q.; and Su, W. J. 2020. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23).

Chen, X.; Chen, T.; Sun, H.; Wu, S. Z.; and Hong, M. 2020. Distributed training with heterogeneous data: Bridging median-and mean-based algorithms. *Advances in Neural Information Processing Systems*, 33: 21616–21626.

Chen, X.; Wu, S. Z.; and Hong, M. 2020. Understanding gradient clipping in private SGD: A geometric perspective. *Advances in Neural Information Processing Systems*, 33: 13773–13782.

Chu, J. T. 1955. On bounds for the normal integral. *Biometrika*, 42(1/2): 263–265.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926. IEEE.

Condat, L.; Yi, K.; and Richtárik, P. 2022. EF-BV: A Unified Theory of Error Feedback and Variance Reduction Mechanisms for Biased and Unbiased Compression in Distributed Optimization. *arXiv preprint arXiv:2205.04180*.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.

Dong, J.; Roth, A.; and Su, W. 2021. Gaussian Differential Privacy. *Journal of the Royal Statistical Society*.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4): 211–407.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33: 16937–16947.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Ghadimi, S.; and Lan, G. 2013. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4): 2341–2368.

Haddadpour, F.; Kamani, M. M.; Mokhtari, A.; and Mahdavi, M. 2021. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2350–2358. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.; Gupta, S.; Song, Z.; Li, K.; and Arora, S. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34: 7232–7241.

Isik, B.; and Weissman, T. 2022. Learning under Storage and Privacy Constraints. *arXiv preprint arXiv:2202.02892*.

Jin, R.; Huang, Y.; He, X.; Dai, H.; and Wu, T. 2020. Stochastic-sign SGD for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*.

Kairouz, P.; Liu, Z.; and Steinke, T. 2021. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, 5201–5212. PMLR.

Kantorovich, L. V.; and Akilov, G. P. 2016. *Functional analysis*. Elsevier.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Karimireddy, S. P.; Rebjock, Q.; Stich, S.; and Jaggi, M. 2019. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, 3252–3261. PMLR.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Krizhevsky, A.; and Hinton, G. 2010. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, Y.; Chang, T.-H.; and Chi, C.-Y. 2020. Secure federated averaging algorithm with differential privacy. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Lian, X.; Zhang, C.; Zhang, H.; Hsieh, C.-J.; Zhang, W.; and Liu, J. 2017. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, 263–275. IEEE.

Mironov, I.; Talwar, K.; and Zhang, L. 2019. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Reisizadeh, A.; Mokhtari, A.; Hassani, H.; Jadbabaie, A.; and Pedarsani, R. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, 2021–2031. PMLR.

Safaryan, M.; and Richtárik, P. 2021. Stochastic sign descent methods: New algorithms and better theory. In *International Conference on Machine Learning*, 9224–9234. PMLR.

Sason, I.; and Verdú, S. 2016. f-divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11): 5973–6006.

Seide, F.; Fu, H.; Droppo, J.; Li, G.; and Yu, D. 2014. 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*. Citeseer.

Suresh, A. T.; Felix, X. Y.; Kumar, S.; and McMahan, H. B. 2017. Distributed mean estimation with limited communication. In *International conference on machine learning*, 3329–3337. PMLR.

Tang, H.; Gan, S.; Awan, A. A.; Rajbhandari, S.; Li, C.; Lian, X.; Liu, J.; Zhang, C.; and He, Y. 2021a. 1-bit adam: Communication efficient large-scale training with adam's convergence speed. In *International Conference on Machine Learning*, 10118–10129. PMLR.

Tang, H.; Yu, C.; Lian, X.; Zhang, T.; and Liu, J. 2019. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, 6155–6165. PMLR.

Tang, Z.; Chang, T.-H.; Ye, X.; and Zha, H. 2021b. Low-rank Matrix Recovery With Unknown Correspondence. *arXiv preprint arXiv:2110.07959*.

Van Erven, T.; and Harremos, P. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7): 3797–3820.

Vargaftik, S.; Ben-Basat, R.; Portnoy, A.; Mendelson, G.; Ben-Itzhak, Y.; and Mitzenmacher, M. 2021. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34: 362–377.

Vogels, T.; Karimireddy, S. P.; and Jaggi, M. 2019. PowerSGD: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32.

Wang, J.; and Joshi, G. 2021. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *Journal of Machine Learning Research*, 22.

Wang, L.; Jia, R.; and Song, D. 2020. D2P-Fed: Differentially private federated learning with efficient communication. *arXiv preprint arXiv:2006.13039*.

Wang, Y.; Xu, Y.; Shi, Q.; and Chang, T.-H. 2021. Quantized federated learning under transmission delay and outage constraints. *IEEE Journal on Selected Areas in Communications*, 40(1): 323–341.

Wu, Y.; and He, K. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yu, H.; Yang, S.; and Zhu, S. 2019. 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, 5693–5700.

Zhang, X.; Chen, X.; Hong, M.; Wu, Z. S.; and Yi, J. 2021. Understanding Clipping for Federated Learning: Convergence and Client-Level Differential Privacy. *arXiv preprint arXiv:2106.13673*.

Zheng, Q.; Chen, S.; Long, Q.; and Su, W. 2021. Federated f-differential privacy. In *International Conference on Artificial Intelligence and Statistics*, 2251–2259. PMLR.

Zheng, S.; Huang, Z.; and Kwok, J. 2019. Communication-efficient distributed blockwise momentum SGD with error-feedback. *Advances in Neural Information Processing Systems*, 32.