Distributed Training with Heterogeneous Data: Bridging Median- and Mean-Based Algorithms

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiangyi Chen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisor: Dr. Mingyi Hong

March, 2022

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school. First and foremost, I would like to express my deepest gratitude to my advisor Prof. Mingyi Hong for providing me great flexibility and guidance in my research. Thanks to him, I am able to work in my interested research areas and finish this dissertation.

Also, I would like to thank my thesis committee members Prof. Mehmet Akcakaya and Prof Jie Ding for for their support and being extremely responsive in scheduling the defense.

The work in this dissertation would not have been possible without the help of my collaborators Haoran Sun, who gave generous help on the experiments, Tiancong Chen, though discussion with whom the key theoretical result was proven, and Prof. Steven Wu, who provided great insights and helped shaping the research direction of this work.

Last but not the least, I would like to thank my family for all their unwavering love and unconditional support in helping me overcome the difficulties encountered.

Abstract

Recently, there is a growing interest in the study of median-based algorithms for distributed non-convex optimization. Two prominent examples include signSGD with majority vote, an effective approach for communication reduction via 1-bit compression on the local gradients, and medianSGD, an algorithm recently proposed to ensure robustness against Byzantine workers. The convergence analyses for these algorithms critically rely on the assumption that all the distributed data are drawn iid from the same distribution. However, in applications such as Federated Learning, the data across different nodes or machines can be inherently heterogeneous, which violates such an iid assumption. This work analyzes signSGD and medianSGD in distributed settings with heterogeneous data. We show that these algorithms are non-convergent whenever there is some disparity between the expected median and mean over the local gradients. To overcome this gap, we provide a novel gradient correction mechanism that perturbs the local gradients with noise, which we show can provably close the gap between mean and median of the gradients. The proposed methods largely preserve nice properties of these median-based algorithms, such as the low per-iteration communication complexity of signSGD, and further enjoy global convergence to stationary solutions. Our perturbation technique can be of independent interest when one wishes to estimate mean through a median estimator.

Contents

List of Figures

Chapter 1

Introduction

In the past few years, deep neural networks have achieved great success in many tasks including computer vision and natural language processing. For many tasks in these fields, it may take weeks or even months to train a model due to the size of the model and training dataset. One practical and promising way to reduce the training time of deep neural networks is using distributed training [Dean et al., 2012]. A popular and practically successful paradigm for distributed training is the parameter server framework [Li et al., 2014], where most of the computation is offloaded to workers in parallel and a parameter sever is used for coordinating the training process. Formally, the goal of such distributed optimization is to minimize the average of M different functions from M nodes,

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{M} \sum_{i=1}^{M} f_i(x), \tag{1.1}$$

where each node i can only access information of its local function $f_i(\cdot)$, defined by its local data. Typically, such local objective takes the form of either a *expected* loss over local data distribution (population risk), or a *empirical average* over loss functions evaluated over a finite number of data points (empirical risk). That is,

$$f_i(x) = \int p_i(\zeta) l(x; \zeta) d\zeta, \quad \text{or} \quad f_i(x) = \frac{1}{K_i} \sum_{k=1}^{K_i} l(x; \zeta_{i,k}) \tag{1.2}$$

where $l(x; \zeta)$ (resp. $l(x; \zeta_{i,k})$) is the cost evaluated at a given data point ζ (resp. $\zeta_{i,k}$).

Similar to the parameter server paradigm, motivated by the use of machine learning models on mobile devices, a distributed training framework called Federated Learning has become popular [Konečný et al., 2016, McMahan et al., 2017, McMahan and Ramage, 2017]. In Federated Learning, the training data are distributed across personal devices and one wants to train a model without transmitting the users' data due to privacy concerns. While many distributed algorithms proposed for parameter server are applicable to Federated Learning, Federated Learning posed many unique challenges, including the presence of *heterogeneous* data across the nodes, and the need to accommodate asynchronous updates, as well as very limited message exchange among the nodes and the servers. By *heterogeneous* data, we mean that either $p_i(\zeta)$, or the empirical distribution formed by $\{\zeta_{i,k}\}_{k=1}^K$ in (1.2), are significantly different across the local nodes. Clearly, when the data is heterogeneous, we will usually have $\nabla f_i(x) \neq \nabla f_j(x)$ and if local data are *homogeneous*, we will have $\nabla f_i(x) = \nabla f_j(x)$ or $\nabla f_i(x) \approx \nabla f_j(x)$ when K is large.

Median-Based Methods. Under these distributed optimization frameworks, many algorithms based on stochastic gradient descent (SGD) have been proposed to solve (1.1). The basic idea is to perform updates based on the *mean* of the local stochastic directions. On the other hand, there are two prominent and interesting algorithms whose updates are *not* directly related to the mean of the local gradients. One is called SIGNSGD (with majority vote) (see Algorithm 1) [Bernstein et al., 2018a], which updates the parameters based on a majority vote of sign of gradients to reduce communication overheads. The other one is called MEDIANGD (see its generalized stochastic version in Algorithm 2, which we refer to as MEDIANSGD [Yin et al., 2018]), which aims to ensure robustness against Byzantine workers by using coordinate-wise median of gradients to evaluate mean of gradients.

Algorithm 1 SIGNSGD (with M nodes)

1: **Input:** learning rate δ, current point x_t
2: $g_{t,i} \leftarrow \nabla f_i(x_t) + \text{sampling noise}$
3: $x_{t+1} \leftarrow x_t - \delta\, \text{sign}(\sum_{i=1}^M \text{sign}(g_{t,i}))$

Algorithm 2 MEDIANSGD (with M nodes)

1: **Input:** learning rate δ, current point x_t
2: $g_{t,i} \leftarrow \nabla f_i(x_t) + \text{sampling noise}$
3: $x_{t+1} \leftarrow x_t - \delta \text{median}(\{g_{t,i}\}_{i=1}^M)$

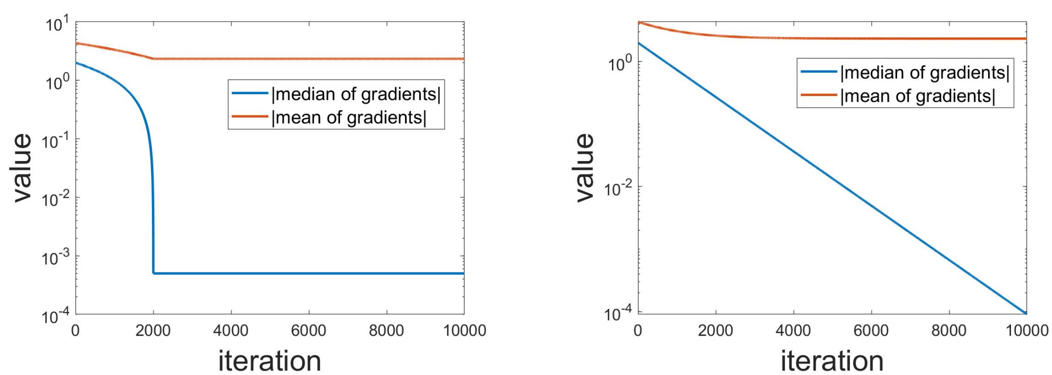
It is clear that SIGNSGD and MEDIANSGD do not simply *average* their local gradients. At first glance, their update rules also appear to be fundamentally different since they are tailored to different desiderata (that is, communication-efficiency versus robustness). Interestingly, in this work we made an observation that, SIGNSGD can be viewed as updating variables along signed median direction (sign(median($\{g_{t,i}\}_{i=1}^M$))), uncovering its hidden connection to MEDIANSGD. This view provides a unified interpretation of these two algorithms as *median-based distributed algorithms*. We analyze these median-based methods in the heterogeneous regime.

Homogeneous v.s. Heterogeneous Data. While the median-based methods are increasingly popular, there has not been a good understanding about the convergence behavior of median-based methods. The existing analyses of both SIGNSGD and MEDIANSGD rely on the assumption of homogeneous data. SIGNSGD is analyzed from the in-sample optimization perspective: it converges to stationary points if the stochastic gradients $g_{t,i}$ sampled from each worker follow the same distribution [Bernstein et al., 2018a,b]. That is, $\nabla f_i(x_t) = \nabla f_j(x_t)$, $\forall$ x_t, and the sampling noises follow the same distribution. On the other hand, MEDIANSGD is analyzed under the framework of population risk minimization: it converges with an optimal statistical rate, but again under the assumption that the data across the workers are iid [Yin et al., 2018].

However, in many modern distributed settings especially Federated Learning, data on different worker nodes can be inherently heterogeneous. For example, users' data stored on different worker nodes might come from different geographic regions, which induce substantially different data distributions. In Federated Learning, the stochastic gradient $g_{t,i}$ from each device is effectively the full gradient $\nabla f_i(x_t)$ evaluated on the user's data (due to the small size of local data), which violates the assumption of identical gradient distributions. Therefore, under these heterogeneous data settings, data aggregation and shuffling are often infeasible, and there is very little understanding of the behavior of both aforementioned algorithms.

From the fixed-point perspective, median-based algorithms like SIGNSGD and MEDIANSGD drive the median of gradients to 0—that is, when the median of gradients reaches 0, the algorithms will not perform updates. When the median is close to the mean of gradients (the latter is the gradient of the target loss function), it follows that the true gradient is also approximately 0, and an approximate stationary solution is reached. The reason for assuming homogeneous data in existing literature [Bernstein et al., 2018a,b, Yin et al., 2018] is exactly to ensure that the median is close to mean. However, when the data from different workers are not drawn from the same distribution, the potential gap between the mean and median could prevent these algorithms from reducing the true gradient.

To illustrate this phenomenon, let us consider a simple one-dimensional example: $\frac{1}{3}\sum_{i=1}^{3} f_i(x) \triangleq (x-a_i)^2/2$, with $a_1 = 1, a_2 = 2, a_3 = 10$. If we run signSGD and medianSGD with step size $\delta = 0.001$ and initial point $x_0 = 0.0005$, both algorithms will produce iterates with large disparity between the mean and median gradients. See Fig. 1.1 for the trajectories of gradient norms. Both algorithms drive the median of gradients to 0 (signSGD finally converges to the level of step size due to the use of sign in its update rule), while the true gradient remains a constant. In Chapter 6, we provide further empirical evaluation to demonstrate that such disparity can severely hinder the training performance.



(a) Trajectory of signSGD (b) Trajectory of medianSGD

Figure 1.1: Absolute value of mean and median of gradient vs iteration. (a) shows the trajectory of signSGD (b) shows the trajectory of medianSGD

Contribution. Motivated by the need to understand median-based algorithms under heterogeneous data settings, we investigate two questions: 1) in a distributed training environment, under what conditions do signSGD and medianSGD work well? and 2) can we provide mechanisms to close the convergence gap in these algorithms? Specifically, we analyze the convergence rate of signSGD and medianSGD, with finite number of data samples per worker node, without assuming the data on different workers are from the same distribution. Our contributions are summarized as follows.

(1) **signSGD as a median-based algorithm**: We show that signSGD updates along the direction of signed *median* of gradients, which connects signSGD and

medianSGD. This fact is crucial for our subsequent analysis of signSGD, and has not been recognized by existing works so far.

(2) **Bridging the gap between median and mean by adding controlled noise.** We prove the following key result: given an arbitrary set of numbers, if one adds unimodal and symmetric noises with variance σ^2, then the expected median of the resulting numbers will approach the expected mean of the original numbers, with a rate of $O(1/\sigma)$. In addition, the distribution of the median will become increasingly symmetric as the variance of noise increases, with a rate of $O(1/\sigma^2)$. This result could be of independent interest.

(3) **Non-convergence due to the gap between median and mean.** We prove that signSGD and medianSGD only converge to solutions whose gradient sizes are proportional to the *difference* between the expected median and mean of gradients at different workers. Further, we show that the non-convergence is not an artifact of analysis by providing examples where signSGD and medianSGD does not converge due to the gap between median and mean.

(4) **Convergence of noisy signSGD and noisy medianSGD.** By using contribution (2), that the expected median converges to mean, and a sharp analysis on the probability density function of the noise on median, we prove that noisy signSGD and noisy medianSGD can both converge to stationary points.

Finally, we emphasize that the connection we established between signSGD and median based method is mainly used to identify and resolve the non-convergence issue of signSGD. The focus of this paper is *not* to analyze properties of median-based methods beyond convergence.

Chapter 2

Preliminaries

2.1 Related work

Distributed training and signSGD. Distributed training of neural nets has become popular since the work of Dean et al. [2012], in which distributed SGD was shown to achieve significant acceleration compared with SGD [Robbins and Monro, 1951]. As an example, Goyal et al. [2017] showed that distributed training of ResNet-50 [He et al., 2016] can finish within an hour. There is a recent line of work providing methods for communication reduction in distributed training, including stochastic quantization [Alistarh et al., 2017, Wen et al., 2017, Wangni et al., 2018] and 1-bit gradient compression such as signSGD [Bernstein et al., 2018a,b]. It is shown in Reddi et al. [2019] that Adam [Kingma and Ba, 2014] can diverge in some cases. Since signSGD is a special case of Adam, it can suffer the same issue as Adam in general and one possible fix to this issue is by using error feedback [Karimireddy et al., 2019]. However, it is shown in Bernstein et al. [2018b] that when noise is unimodal and symmetric, signSGD can guarantee convergence.

Byzantine robust optimization. Byzantine robust optimization draws increasingly more attention in the past few years. Its goal is to ensure the performance of the optimization algorithms in the existence of Byzantine failures. Alistarh et al. [2018] developed a variant of SGD based on detecting Byzantine nodes. Yin et al. [2018] proposed medianGD that is shown to converge with optimal statistical rate. Blanchard et al. [2017] proposed a robust aggregation rule called Krum. It is shown in Bernstein et al. [2018b] that SIGNSGD is also robust against certain failures. Most existing works assume homogeneous data. In addition, Bagdasaryan et al. [2018] showed that many existing Byzantine robust methods are vulnerable to adversarial attacks.

Federated Learning. Federated Learning was initially introduced in Konečnỳ et al. [2016], McMahan and Ramage [2017] for collaborative training of machine learning models without transmitting users' data. It is featured by high communication costs, requirements for failure tolerance and privacy protection, as the nodes are likely to be mobile devices such as cell phones. Smith et al. [2017] proposed a learning framework that incorporates multi-task learning into Federated Learning. Bonawitz et al. [2019] proposed system design for large scale Federated Learning. There is a line of work on design and analysis of algorithms in federated learning includes Sattler et al. [2019], Reisizadeh et al. [2019], Zhou and Cong [2017], Stich [2018].

2.2 Notations

Given a set of vectors a_i, $i = 1, ..., n$, we denote $\{a_i\}_{i=1}^n$ to be the the set and denote $\text{median}(\{a_i\}_{i=1}^n)$ to be the coordinate-wise median of of the vectors. We also use $\text{median}(\{a\})$ to denote $\text{median}(\{a_i\}_{i=1}^n)$ for simplicity. When v is a vector and b is a constant, $v \neq b$ means none of the coordinate of v equals b. Finally, $(v)_j$ denotes jth coordinate of v, $\text{sign}(v)$ denotes the signed vector of v. We use $[N]$ to denote the set $\{1, 2, \cdots, N\}$.

Chapter 3

Distributed signSGD and medianSGD

In this chapter, we give convergence analyses of signSGD and medianSGD for the problem defined in (1.1), without any assumption on data distribution. We first analyze the convergence of the algorithms under the framework of stochastic optimization. In such a setting, at iteration t, worker i can access a stochastic gradient estimator $\hat{g}_i(x_t)$ (also denoted as $\hat{g}_{t,i}$ for simplicity). Denote the collection of the stochastic gradients to be $\{\hat{g}_t\}$. we make following assumptions thoughout the paper. A1. Unbiased gradient estimator, $\mathbb{E}[g_i(x)] = \nabla f_i(x)$. A2. Bounded variance, $\mathbb{E}[\|\text{median}(\{g_t\})_j - \mathbb{E}[\text{median}(\{g_t\})_j|x_t]\|^2] \leq \sigma_m^2,\ \forall j \in [d]$. A3. f has Lipschitz gradient, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. A4. M is an odd number. The assumptions A1 and A3 are standard for stochastic optimization. A4 ensures median is well defined. A2 is a variant of classical bounded variance assumption for median-based algorithms and it is satisfied if all $\hat{g}_{t,i}$ has bounded variance.

3.1 Convergence of signSGD and medianSGD

From the pseudo-code of Algorithm 1, it is not straightforward to see how signSGD is related to the median of gradients, since there is no explicit operation for calculating median in the update rule of signSGD. It turns out that signSGD actually goes along the signed median direction.

Proposition 1. *When M is odd and* $\text{median}(\{g_t\}) \neq 0$*, we have*

$$\text{sign}(\sum_{i=1}^{M} \text{sign}(g_{t,i})) = \text{sign}(\text{median}(\{g_t\})). \tag{3.1}$$

Proof: Suppose we have a set of numbers $a_k, k \in [M]$, $a_k \neq 0, \forall k$ and M is odd. We show the following identity

$$\text{sign}\left(\sum_{k=1}^{M} \text{sign}(a_k)\right) = \text{sign}(\text{median}(\{a_k\}_{k=1}^{M})) \tag{3.2}$$

To begin with, define $b_k, k \in [M]$ to be a sequence of a_k sorted in ascending order. Then we have

$$\text{median}(\{a_k\}_{k=1}^{M}) = \text{median}(\{b_k\}_{k=1}^{M}) = b_{(M+1)/2} \tag{3.3}$$

and the following

$$\begin{aligned}\text{sign}\left(\sum_{k=1}^{M} \text{sign}(a_k)\right) =& \text{sign}\left(\sum_{k=1}^{M} \text{sign}(b_k)\right) \\ =& \text{sign}\left(\text{sign}(b_{(M+1)/2}) + \sum_{k=1}^{(M+1)/2-1} \text{sign}(b_k) + \sum_{k=(M+1)/2+1}^{M} \text{sign}(b_k)\right).\end{aligned} \tag{3.4}$$

Recall that b_k is non-decreasing as it is a sorted sequence of a_k with ascending order. If $b_{(M+1)/2} > 0$, we have $b_k > 0, \forall k > (M+1)/2$ and thus

$$\sum_{k=(M+1)/2+1}^{M} \text{sign}(b_k) = \sum_{k=(M+1)/2+1}^{M} 1 = (M-1)/2. \tag{3.5}$$

Since $\sum_{k=(M+1)/2+1}^{M} \text{sign}(b_k) \geq \sum_{k=(M+1)/2+1}^{M} -1 = -(M-1)/2$, we have

$$\text{sign}(b_{(M+1)/2}) + \sum_{k=1}^{(M+1)/2-1} \text{sign}(b_k) + \sum_{k=(M+1)/2+1}^{M} \text{sign}(b_k) \geq (\text{ sign}(b_k)) = 1 \tag{3.6}$$

which means when $\text{median}(a_k) > 0$,

$$\text{sign}\left(\sum_{k=1}^{M} \text{sign}(a_k)\right) = 1 \tag{3.7}$$

Following the same procedures as above, one can also get when $\text{median}(a_k) < 0$,

$$\text{sign}\left(\sum_{k=1}^{M} \text{sign}(a_k)\right) = -1 \tag{3.8}$$

Thus,

$$\text{sign}\left(\sum_{k=1}^{M} \text{sign}(a_k)\right) = \text{sign}\left(\text{median}(a_k)\right) \tag{3.9}$$

when $\text{median}(a_k) \neq 0$.

Applying the result above to each coordinate of the gradient vectors finishes the proof.

Thus, signSGD updates the variables based on the sign of coordinate-wise median of gradients, while medianSGD updates the variables toward the median direction of gradients. Though these two algorithms are closely related, their convergence behaviors are not well-understood. We provide the convergence guarantee for these algorithms in Theorem 1 and Theorem 2, respectively.

Theorem 1. *Suppose A1-A4 are satisfied, and define* $D_f \triangleq f(x_1) - \min_x f(x)$. *For* signSGD *with* $\delta = \frac{\sqrt{D_f}}{\sqrt{LdT}}$, *the following holds true*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(x_t)\|_1] \leq \frac{3}{2}\frac{\sqrt{dLD_f}}{\sqrt{T}} + 2d\sigma_m + 2\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]|x_t] - \nabla f(x_t)\|_1]. \tag{3.10}$$

Proof: Let us define:

$$\text{median}(\{g_t\}) \triangleq \text{median}(\{g_{t,i}\}_{i=1}^{M}). \tag{3.11}$$

and

$$\text{median}(\{\nabla f_t\}) \triangleq \text{median}(\{\nabla f_i(x_t)\}_{i=1}^{M}). \tag{3.12}$$

By A3, we have the following standard descent lemma in nonconvex optimization.

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \tag{3.13}$$

Substituting the update rule into (3.13), we have the following series of inequalities

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \delta\langle\nabla f(x_t), \text{sign}(\text{median}(\{g_t\}))\rangle + \frac{L}{2}\delta^2 d \\ &= f(x_t) - \delta\langle\mathbb{E}[\text{median}(\{g_t\})], \text{sign}(\text{median}(\{g_t\}))\rangle \\ &\quad + \delta\langle\mathbb{E}[\text{median}(\{g_t\})] - \nabla f(x_t), \text{sign}(\text{median}(\{g_t\}))\rangle + \frac{L}{2}\delta^2 d \\ &\leq f(x_t) - \delta\|\mathbb{E}[\text{median}(\{g_t\})]\|_1 + \delta\|\mathbb{E}[\text{median}(\{g_t\})] - \nabla f(x_t)\|_1 \\ &+ 2\delta\sum_{j=1}^{d}|\mathbb{E}[\text{median}(\{g_t\})_j]| I[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\mathbb{E}[\text{median}(\{g_t\})_j])] \\ &+ \frac{L}{2}\delta^2 d \end{aligned} \tag{3.14}$$

where $\text{median}(\{g_t\})_j$ is jth coodrinate of $\text{median}(\{g_t\})$, and $I[\cdot]$ denotes the indicator function.

Taking expectation over all the randomness, we get

$$\begin{aligned} &\delta\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]\|_1 \\ &\leq \mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})] + \delta\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] \\ &+ 2\delta E\left[\sum_{j=1}^{d}|\mathbb{E}[\text{median}(\{g_t\})_j]| P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\mathbb{E}[\text{median}(\{g_t\})_j])]\right] \\ &+ \frac{L}{2}\delta^2 d \end{aligned} \tag{3.15}$$

Before we proceed, we analyze the error probability of sign

$$P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\mathbb{E}[\text{median}(\{g_t\})_j])] \tag{3.16}$$

This follows a similar analysis as in SIGNSGD paper.

By reparameterization, we can have

$$\text{median}(\{g_t\})_j = \mathbb{E}[\text{median}(\{g_t\})_j] + \zeta_{t,j}$$

with $\mathbb{E}[\zeta_{t,j}] = 0$.

By Markov inequality and Jensen's inequality, we have

$$\begin{aligned}
&P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\mathbb{E}[\text{median}(\{g_t\})_j])] \\
&\leq P[|\zeta_{t,j}| \geq \mathbb{E}[\text{median}(\{g_t\})_j]] \\
&\leq \frac{\mathbb{E}[|\zeta_{t,j}|]}{\mathbb{E}[\text{median}(\{g_t\})_j]} \\
&\leq \frac{\sqrt{\mathbb{E}[\zeta_{t,j}^2]}}{\mathbb{E}[\text{median}(\{g_t\})_j]} = \frac{\sigma_m}{\mathbb{E}[\text{median}(\{g_t\})_j]}
\end{aligned} \tag{3.17}$$

where we assumed $\mathbb{E}[\zeta_{t,j}^2] \leq \sigma_m^2$.

Substitute (3.17) into (3.15), we get

$$\begin{aligned}
&\delta\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]\|_1] \\
&\leq \mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})] + \delta\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] + 2\delta d\sigma_m + \frac{L}{2}\delta^2 d.
\end{aligned} \tag{3.18}$$

Now we use standard approach to analyze convergence rate. Summing over t from 1 to T and divide both sides by $T\delta$, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]\|_1] \\
&\leq \frac{D_f}{T\delta} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] + 2d\sigma_m + \frac{L}{2}\delta d
\end{aligned} \tag{3.19}$$

where here we defined $D_f \triangleq \mathbb{E}[f(x_1)] \min_x f(x)$

Now set $\delta = \frac{\sqrt{D_f}}{\sqrt{LdT}}$, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]\|_1] \\
&\leq \frac{3}{2}\frac{\sqrt{dLD_f}}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] + 2d\sigma_m
\end{aligned} \tag{3.20}$$

Going one step further, and use the triangular inequality, we can easily bound the ℓ_1 norm of the gradient as the following

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|_1] \leq \frac{3}{2}\frac{\sqrt{dLD_f}}{\sqrt{T}} + 2\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] + 2d\sigma_m \tag{3.21}$$

One key observation from Theorem 1 is that as T goes to infinity, the RHS of (3.10) is dominated by the difference between the median and mean and the standard deviation on the median.

We remark that under the assumption that the gradient estimators from different nodes are drawn from the same unimodal and symmetric distribution in signSGD, the analysis recovers the bound in Bernstein et al. [2018a]. In this case, we have $\mathbb{E}[\text{median}(\{g_t\})|x_t] = \mathbb{E}[\nabla f(x_t)]$ and $\sigma_m = O(\sigma_l/\sqrt{M})$ if the noise on each coordinate of local gradients has variance bounded by σ_l^2 (see Theorem 1.4.1 in Miller [2017]). Then, (3.10) becomes $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)]\|_1 \leq \frac{3}{2}\frac{\sqrt{dLD_f}}{\sqrt{T}} + dO(\frac{\sigma_l}{\sqrt{M}})$. Under minibatch setting, one can further use a large minibatch to decrease σ_l to show better convergence.

Theorem 2. *Suppose A1-A4 are satisfied, define* $D_f \triangleq f(x_1) - \min_x f(x)$. *Set* $\delta = \min(\frac{1}{\sqrt{Td}}, \frac{1}{2L})$, medianSGD *yields*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2\sqrt{d}}{\sqrt{T}}D_f + 2L\frac{\sqrt{d}}{\sqrt{T}}\sigma_m^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2]. \tag{3.22}$$

Proof: By the gradient Lipschitz continuity and the update rule, we have

$$\begin{aligned}
& f(x_{t+1}) \\
\leq & f(x_t) - \delta\langle\nabla f(x_t), \text{median}(\{g_t\})\rangle + \frac{L}{2}\delta^2\|\text{median}(\{g_t\})\|^2 \\
\leq & f(x_t) - \delta\langle\nabla f(x_t), \text{median}(\{g_t\})\rangle \\
& + L\delta^2(\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2 + \|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2)
\end{aligned}$$

Taking expectation, we have

$$
\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - \mathbb{E}[f(x_t)] \\
\leq &- \delta \mathbb{E}_{x_t}[\langle \nabla f(x_t), \mathbb{E}_{\{g_t\}}[\text{median}(\{g_t\})|x_t]\rangle] \\
&+ L\delta^2 \mathbb{E}[\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2 + \|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
= &- \delta \mathbb{E}_{x_t}\left[\frac{1}{2}\left(\|\nabla f(x_t)\|^2 + \|\mathbb{E}_{\{g_t\}}[\text{median}(\{g_t\})|x_t]\|^2\right)\right] \\
&+ \delta \mathbb{E}_{x_t}\left[\frac{1}{2}\|\nabla f(x_t) - \mathbb{E}_{\{g_t\}}[\text{median}(\{g_t\})|x_t]\|^2\right] \\
&+ L\delta^2 \mathbb{E}[\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2 + \|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
= &- \frac{\delta}{2}\mathbb{E}\left[\|\nabla f(x_t)\|^2 + \|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2 - \|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2\right] \\
&+ L\delta^2 \mathbb{E}[\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2 + \|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
= &- \frac{\delta}{2}\mathbb{E}[\|\nabla f(x_t)\|^2] - (\frac{\delta}{2} - L\delta^2)\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
&+ \frac{\delta}{2}\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
&+ L\delta^2 \mathbb{E}[\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2]
\end{aligned} \tag{3.23}
$$

where $\mathbb{E}_{x_t}[\cdot]$ is expectation over randomness of x_t and $\mathbb{E}_{\{g_t\}}[\cdot|x_t]$ is expectation over randomness of $\{g_t\}$ given x_t.

Setting $\delta = \min(\frac{1}{\sqrt{Td}}, \frac{1}{2L})$, telescope sum and divide both sides by $T\delta/2$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \\
\leq &\frac{2\sqrt{d}}{\sqrt{T}}(\mathbb{E}[f(x_1)] - \mathbb{E}[f(x_{T+1})]) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] + 2L\frac{\sqrt{d}}{\sqrt{T}}\sigma_m^2
\end{aligned} \tag{3.24}
$$

Substituting $\mathbb{E}[f(x_1)] - \mathbb{E}[f(x_{T+1})] \leq D_f$ into the above inequality completes the proof.

As T increases, the RHS of (3.22) will be dominated by the terms involving the difference between the expected median of gradients and the true gradients. In the case where the gradient from each node follows the same symmetric and unimodal distribution, the difference vanishes and the algorithm converges to a stationary point with a rate of $\frac{\sqrt{d}}{\sqrt{T}}$. However, when the gap between the expected median of gradients and the true gradients is not zero, our results suggest that both signSGD and medianSGD can only converge to solutions where the size of the gradient is upper bounded by some constant related to the median-mean gap.

3.2 Tightness of the convergence analysis

Theorem 1 – 2 suggest that it is difficult to make the *upper bounds* on the average size of the gradient of signSGD and medianSGD go to zero. We now provide examples to demonstrate that such a convergence gap indeed exists, thus showing that the gap in the convergence analysis is inevitable unless additional assumptions are enforced.

Theorem 3. *There exists a problem instance where* signSGD *converges to a point* $\hat{x}^*$ *with*

$$\|\nabla f(\hat{x}^*)\|_1 \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] \geq 1$$

and medianSGD *converges to*

$$\|\nabla f(\hat{x}^*)\|^2 \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|^2] \geq 1.$$

Proof: We show that our analysis is tight, in the sense that the constant gap

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] \tag{3.25}$$

does exist in practice.

Consider the following problem

$$\min_{x\in\mathbb{R}} f(x) \triangleq \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}(x-a_i)^2 \tag{3.26}$$

with $a_1 < a_2 < a_3$. In particular, $f_i(x) = \frac{1}{2}(x-a_i)^2$, so each local node has only one data point. Since the entire problem is deterministic, and the local gradient is also deterministic (i.e., no subsampling is available), we will drop the expectation below.

It is readily seen that the median of gradient is always $x - a_2$. Therefore running signSGD on the above problem is equivalent to running signSGD to minimize

$\frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}(x-a_2)^2$. From the Theorem 1 in Bernstein et al. [2018a], the SIGNSGD will converge to $x=a_2$ as T goes to ∞ and $\delta = O(\frac{1}{\sqrt{T}}))$.

On the other hand, at the point $x=a_2$, the median of gradients median($\{g_t\}$) is 0 but the gradient of $f(x)$ is given by

$$\nabla f(a_2) = \frac{1}{3}\sum_{i=1}^{3}(x-a_i) = \frac{1}{3}((a_2-a_3)+(a_2-a_1)) \tag{3.27}$$

Recall that for this problem, we also have for any x_t,

$$\begin{aligned} &\|\mathbb{E}[\text{median}(\{g_t\})] - \nabla f(x_t)\|_1 \\ &= \|\text{median}(\{g_t\}) - \nabla f(x_t)\|_1 \\ &= \left| x_t - a_2 - \frac{1}{3}\sum_{i=1}^{3}(x_t-a_i)\right| = \left|\frac{1}{3}(2a_2-a_1-a_3)\right|. \end{aligned} \tag{3.28}$$

Comparing (3.27) and (3.28), we conclude that at a given point $x=a_2$ (for which the SIGNSGD will converge to), we have

$$\|\nabla f(x)\|_1 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|_1] = \left|\frac{1}{3}(2a_2-a_1-a_3)\right|. \tag{3.29}$$

Substituting $a_1=0, a_2=1, a_3=5$ (which satisfies $a_1<a_2<a_3$ assumed the beginning) into (3.29) finishes the proof for SIGNSGD.

The proof for MEDIANSGD uses the same construction as the proof of SIGNSGD, i.e. we consider the problem

$$\min_{x\in\mathbb{R}} f(x) \triangleq \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}(x-a_i)^2 \tag{3.30}$$

with $a_1<a_2<a_3$. Then from the update rule of MEDIANSGD, it reduces to running gradient descent to minimize $\frac{1}{2}(x-a_2)^2$. From classical results on convergence of gradient descent, the algorithm will converge to $x=a_2$ with any stepsize $\delta<2/L$.

At the point $x=a_2$, the median of gradients is zero but $\nabla f(x)$ is

$$\nabla f(a_2) = \frac{1}{3}\sum_{i=1}^{3}(x-a_i) = \frac{1}{3}((a_2-a_3)+(a_2-a_1)). \tag{3.31}$$

In addition, for any x_t, the gap between median and mean of gradients satisfy

$$\begin{aligned}&\|\mathbb{E}[\text{median}(\{g_t\})] - \nabla f(x_t)\|^2 \\ =& \left| x_t - a_2 - \frac{1}{3}\sum_{i=1}^{3}(x_t - a_i) \right|^2 = \left| \frac{1}{3}(2a_2 - a_1 - a_3) \right|^2 \end{aligned} \tag{3.32}$$

Combining all above, we have for $x = a_2$, we get

$$\|\nabla f(x)\|^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})]] - \nabla f(x_t)\|^2] = \left| \frac{1}{3}(2a_2 - a_1 - a_3) \right|^2. \tag{3.33}$$

Setting $a_1 = 0, a_2 = 1, a_3 = 5$ we get $|\frac{1}{3}(2a_2 - a_1 - a_3)|^2 = 1$ and the proof is finished.

We remark that the counter example is similar to the example for Figure 1.1, where the divergence is caused by the difference between median and mean. It is shown in Karimireddy et al. [2019] that SIGNSGD with $M = 1$ can diverge due to data sampling but we show SIGNSGD with $M \geq 1$ can diverge even if data sampling is not used. The possible divergence of MEDIANSGD is more or less noticed by the authors when developing median-based algorithms such Yin et al. [2018], Alistarh et al. [2018]. Yet, these works circumvent this issue by assuming iid data. We emphasize this non-convergence issue here because we assume non-iid data.

A traditional way to create iid data distribution in distributed training is to aggregate and shuffle the data. However, in settings with sensitive or private data such as Federated Learning, data shuffling is prohibited. This poses the question that whether it is possible to improve the performance of these median-based algorithms without transmitting data. In the following, we provide a data agnostic method to decrease the gap between median and mean, which will be used to improve median-based algorithms later.

Chapter 4

Convergence of median towards the mean

In the previous chapter, we saw that there could be a convergence gap depending on the difference between expected median and mean for either signSGD or medianSGD. In the following, we present a general result showing that the expected median and the mean can be closer to each other if some random noise is properly added. This is the key leading to our perturbation mechanism to be proposed shortly, which ensures that signSGD and medianSGD can converge properly.

Theorem 4. *Assume we have a set of numbers $u_1, .., u_{2n+1}$. Given a symmetric and unimodal noise distribution with mean 0, variance 1. Denote the pdf of the distribution to be $h_0(z)$ and cdf to be $H_0(z)$. Suppose $h_0'(z)$ is uniformly bounded and absolutely integrable. Draw $2n+1$ samples $\xi_1, ..., \xi_{2n+1}$ from the distribution $h_0(z)$. Define random variable $\hat{u}_i = u_i + b\xi_i$ and $\bar{u} \triangleq \sum_{i=1}^{2n+1} u_i$,*

(a) We have

$$\mathbb{E}[\text{median}(\{\hat{u}_i\}_{i=1}^{2n+1})] = \bar{u} + O\left(\frac{\max_{i,j}|u_i - u_j|^2}{b}\right), \tag{4.1}$$

$$\text{Var}(\text{median}(\{\hat{u}_i\}_{i=1}^{2n+1})) = O(b^2). \tag{4.2}$$

(b) Further assume $h_0''(z)$ is uniformly bounded and absolutely integrable. Denote

$r_b(z)$ to be the pdf of the distribution of $\text{median}(\{\hat{u}_i\}_{i=1}^{2n+1})$*, we have*

$$r_b(\bar{u}+z) = \underbrace{\frac{1}{b}g(\frac{z}{b})}_{\text{symmetric part}} + \underbrace{\frac{1}{b}v(\frac{z}{b})}_{\text{asymmetric part}} \tag{4.3}$$

where

$$g(z) \triangleq \sum_{i=1}^{2n+1} h_0(z) \sum_{S\in\mathcal{S}_i} \prod_{j\in S} H_0(z) \prod_{k\in[n]\setminus\{i,S\}} H_0(-z) \tag{4.4}$$

being the pdf of sample median of $2n+1$ samples drawn from the distribution $h_0(z)$ which is symmetric over 0, $\mathcal{S}_i$ is the set of all n-combinations of items from the set $[2n+1]\setminus i$, and the asymmetric part satisfies

$$\int_{-\infty}^{\infty} \frac{1}{b}|v(\frac{z}{b})|dz = O\left(\frac{\max_i |\bar{u}-u_i|^2}{b^2}\right). \tag{4.5}$$

Proof:

Proof for (a):

Assume we have a set of numbers $u_1, .., u_{2n+1}$. Given a symmetric and unimodal noise distribution with mean 0 and variance 1, denote its pdf to be $h_0(z)$ and its cdf to be $H_0(z)$. Draw $2n+1$ samples from the distribution $\xi_1, ..., \xi_{2n+1}$.

Given a constant b, define random variable $\hat{u}_i = u_i + b\xi_i$. Define $\tilde{u} \triangleq \text{median}(\{\hat{u}_i\}_{i=1}^{2n+1})$ and its pdf and cdf to be $h(z)$ and $H(z)$, respectively. Define $\bar{u} \triangleq \frac{1}{2n+1}\sum_{i=1}^{2n+1} u_i$.

Denote the pdf and cdf of $\hat{u}_i$ to be $h_i(z,b)$ and $H_i(z,b)$. Since $\hat{u}_i = u_i + b\xi_i$ is a scaled and shifted version of ξ_i, given ξ_i has pdf $h_0(z)$ and cdf $H_0(z)$, we know $h_i(z,b) = \frac{1}{b}h_0(\frac{z-u_i}{b})$ and $H_i(z,b) = H_0(\frac{z-u_i}{b})$ from basic probability theory. In addition, from symmetricity of $h_0(z)$, we also have $1 - H_0(z) = H_0(-z)$.

Define pdf of $\tilde{u}$ to be $h(z,b)$, from order statistic, we know

$$h(z,b) = \sum_{i=1}^{2n+1} h_i(z,b) \sum_{S\in\mathcal{S}_i} \prod_{j\in S} H_j(z,b) \prod_{k\in[2n+1]\setminus\{i,S\}} (1 - H_k(z,b)) \tag{4.6}$$

where $\mathcal{S}_i$ is the set of all n-combinations of items from the set $[2n+1]\setminus i$.

To simplify notation, we write the pdf into a more compact form

$$h(z,b) = \sum_{i,\{J,K\}\in\mathcal{S}'_i} h_i(z,b) \prod_{j\in J} H_j(z,b) \prod_{k\in K} (1 - H_k(z,b)) \tag{4.7}$$

where the set $\mathcal{S}_i'$ is the set of all possible $\{J, K\}$ with J being a combination of n items from $[2n+1] \setminus i$ and $K = [2n+1] \setminus \{J, i\}$ and $i \in [2n+1]$ is omitted.

Then the expectation of median can be calculated as

$$
\begin{aligned}
&\mathbb{E}[\tilde{u}] \\
=&\int_{-\infty}^{\infty} z \sum_{i,\{J,K\}\in\mathcal{S}_i'} h_i(z,b) \prod_{j\in J} H_j(z,b) \prod_{k\in K} (1 - H_k(z,b))dz \\
=&\sum_{i,\{J,K\}\in\mathcal{S}_i'} \int_{-\infty}^{+\infty} (bz + u_i)\frac{1}{b}h_0(z) \prod_{j\in J} H_0(z + \frac{u_i - u_j}{b}) \prod_{k\in K} (1 - H_0(z + \frac{u_i - u_k}{b}))bdz \\
=&\sum_{i,\{J,K\}\in\mathcal{S}_i'} \int_{-\infty}^{+\infty} (bz + u_i)h_0(z) \\
&\prod_{j\in J} \left(H_0(z) + \frac{u_i - u_j}{b} h_0(z) + \frac{(u_i - u_j)^2}{2b^2} h_0'(z_j') \right) \\
&\prod_{k\in K} \left(1 - H_0(z) - \frac{u_i - u_k}{b} h_0(z) - \frac{(u_i - u_j)^2}{2b^2} h_0'(z_k') \right) dz
\end{aligned}
$$

where the second inequality is due to a changed of variable from z to $\frac{z-u_i}{b}$, the last inequality is due to Taylor expansion and $z_j' \in [z_j, z_j + \frac{u_i-u_j}{\sigma}]$, $z_k' \in [z_k, z_k + \frac{u_i-u_k}{\sigma}]$.

Now we consider terms with different order w.r.t b after expanding the Taylor expansion.

First, we start with the terms that is multiplied by b, the summation of coefficients in front of these terms equals

$$
\sum_{i,\{J,K\}\in\mathcal{S}_i'} \int_{-\infty}^{+\infty} zh_0(z) \prod_{j\in J} H_0(z)^n \prod_{k\in K} (1 - H_0(z))^n dz = 0
$$

due to symmetricity of f over 0.

Then we consider the terms that are not multiplied by b, the summation of their coefficients equals

$$
\begin{aligned}
&\sum_{i,\{J,K\}\in\mathcal{S}_i'}(u_i-u_j)(\int_{-\infty}^{+\infty} zh_0(z)\prod_{j\in J}H_0(z)^{n-1}\prod_{k\in K}(1-H_0(z))^n h_0(z)dz)\\
&-\sum_{i,\{J,K\}\in\mathcal{S}_i'}(u_i-u_j)(\int_{-\infty}^{+\infty} zh_0(z)\prod_{j\in J}H_0(z)^{n}\prod_{k\in K}(1-H_0(z))^{n-1} h_0(z)dz)\\
&+\sum_{i,\{J,K\}\in\mathcal{S}_i'}u_i(\int_{-\infty}^{+\infty} h_0(z)H_0(z)^n(1-H_0(z))^n dz)\\
=&0+0+\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\int_{-\infty}^{+\infty}H_0(z)^n(1-H_0(z))^n dH_0(z)
\end{aligned}
$$

due to the cancelling in the summation (i.e. $\sum_{i,\{J,K\}\in\mathcal{S}_i'}(u_i-u_j)=0$).

Further, we have

$$
\begin{aligned}
&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\int_{-\infty}^{+\infty}H_0(z)^n(1-H_0(z))^n dH_0(z)\\
\overset{(a)}{=}&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\int_0^1 y^n(1-y)^n dy\\
=&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\frac{1}{n+1}\int_0^1(1-y)^n dy^{n+1}\\
=&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\frac{1}{n+1}\left(-\int_0^1 y^{n+1}d(1-y)^n\right)\\
\overset{(b)}{=}&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\frac{n}{n+1}\int_0^1 y^{n+1}(1-y)^{n-1}dy\\
=&\cdots\\
=&\sum_{i=1}^{2n+1}u_i\binom{2n}{n}\frac{n(n-1)\cdots 1}{(n+1)(n+2)\cdots 2n}\int_0^1 y^{2n}dy\\
=&\sum_{i=1}^{2n}u_i\binom{2n}{n}\frac{n!n!}{(2n+1)!}=\frac{1}{2n+1}\sum_{i=1}^{2n+1}u_i
\end{aligned}
$$

where (a) is due to a change of variable from $H_0(z)$ to y and the omitted steps are just repeating steps from (a) to (b).

In the last step, we consider the rest of the terms (terms multiplied by $1/b$ or higher order w.r.t. $1/b$). Since h_0, h_0' are bounded, for any non-negative integer p, q, k, there exists a constant $c > 0$ such that:

$$\begin{aligned}
&\left|\int_{-\infty}^{+\infty} z h_0(z)(H_0(z)^p h_0(z)^q h_0'(z')^k) dz\right| \\
\leq &\int_{-\infty}^{+\infty} |z||h_0(z)||(H_0(z)^p h_0(z)^q h_0'(z')^k)| dz \\
&\leq c\int_{-\infty}^{+\infty} |z| h_0(z) dz \\
&= c(\int_{-1}^{+1} |z| h_0(z) dz + \int_{1}^{+\infty} |z| h_0(z) dz + \int_{-\infty}^{-1} |z| h_0(z) dz) \\
&\leq c(\int_{-1}^{+1} h_0(z) dz + \int_{1}^{+\infty} z^2 h_0(z) dz + \int_{-\infty}^{-1} z^2 h_0(z) dz) \\
&\leq c(\int_{-1}^{+1} h_0(z) dz + \int_{-\infty}^{+\infty} z^2 h_0(z) dz) \\
&\leq c(\int_{-1}^{+1} h_0(z) dz + 1) \\
&\leq c \qquad \text{[Here's another constant still denoted as } c\text{]}
\end{aligned}$$

And also

$$\begin{aligned}
\left|\int_{-\infty}^{+\infty} h_0(z)(H_0(z)^p h_0(z)^q h_0'(z')^k) dz\right| \leq &\int_{-\infty}^{+\infty} h_0(z)|H_0(z)^p h_0(z)^q h_0'(z')^k| dz \\
\leq &c' \int_{-\infty}^{+\infty} h_0(z) dz = c'
\end{aligned}$$

for some constant c'.

Then the coefficient of rest of the terms are bounded by constant, and the order of them are at least $\mathcal{O}(\frac{1}{b})$. Therefore $|\mathbb{E}[\text{median}(\{\hat{u}_i\}_{i=1}^{2n+1}) - \frac{1}{2n+1}\sum_{i=1}^{2n+1} u_i]| = \mathcal{O}(\frac{1}{b})$ which proves (4.1).

Now we compute the order of the variance of median($\hat{u}_i$) in terms of b

$$
\begin{aligned}
&\text{Var}(\text{median}(\hat{u}_i)) \\
=&\mathbb{E}[\text{median}(\hat{u}_i)^2] - \mathbb{E}[\text{median}(\hat{u}_i)]^2 \\
\leq&\mathbb{E}[\text{median}(\hat{u}_i)^2] \\
=&\sum_{i,\{J,K\}\in\mathcal{S}'_i} \int_{-\infty}^{+\infty} z^2 h_i(z) \prod_{j\in J} H_j(z) \prod_{k\in K} (1 - H_k(z)) dz \\
=&\sum_{i,\{J,K\}\in\mathcal{S}_i} \int_{-\infty}^{+\infty} (bz+u_i)^2 h_0(z) \prod_{j\in J} H_0(z + \frac{u_i - u_j}{b}) \prod_{k\in K} (1 - H_0(z + \frac{u_i - u_k}{b}))dz \\
=&\sum_{i,\{J,K\}\in\mathcal{S}_i} \int_{-\infty}^{+\infty} (bz+u_i)^2 h_0(z) \times \\
&\prod_{j\in J} \left(H_0(z) + \frac{u_i - u_j}{b} h_0(z) + \frac{(u_i-u_j)^2}{2b^2} h_0'(z_j') \right) \\
&\prod_{k\in K} \left(1 - H_0(z) - \frac{u_i - u_k}{b} h_0(z) - \frac{(u_i-u_j)^2}{b^2} h_0'(z_k') \right) dz
\end{aligned}
$$

where $z_j' \in [z_j, z_j + \frac{u_i - u_j}{b}], z_k' \in [z_k, z_k + \frac{u_i - u_k}{b}]$. Similar to the analysis in computing order of gap between median and mean, we consider terms after expanding the multiple formula. Note that we similarly have:

$$
\begin{aligned}
\left| \int_{-\infty}^{+\infty} z^2 h_0(z) (H_0(z)^p h_0(z)^q h_0'(z')^k) dz \right| &\leq \int_{-\infty}^{+\infty} z^2 h_0(z) |(H_0(z)^p h_0(z)^q h_0'(z')^k)| dz \\
&\leq c \int_{-\infty}^{+\infty} z^2 h_0(z) dz \\
&= c
\end{aligned}
$$

Therefore, after expansion and integration, the coefficients of any order of b are also bounded by constant. Since the order of the terms w.r.t b are less than 2, we can conclude that the variance of Median($\hat{u}_i$) is of order $\mathcal{O}(b^2)$ which proves (4.2).

Proof for (b):

This key idea of the proof in part is similar to that for part (a). We use Taylor expansion to expand different terms in pdf of sample median and identify the coefficient in front terms with different order w.r.t. b. The difference is that instead of doing second order Taylor expansion on H_0, we also need to do it for h_0, thus requiring h_0'' to be uniformly bounded and absolutely integrable. In addition, not every higher order term is multiplied by $h_0(z)$, thus more efforts are required for bounding the integration of higher order terms.

First, by a change of variable (change z to $\frac{z-\bar{u}}{b}$), (4.6) can be written as

$$\begin{aligned} &h(\bar{u}+bz,b) \\ &=\sum_{i=1}^{2n+1}\frac{1}{b}h_0(\frac{\bar{u}-u_i}{b}+z)\sum_{S\in\mathcal{S}_i}\prod_{j\in S}H_0(\frac{\bar{u}-u_j}{b}+z)\prod_{k\in[2n+1]\setminus\{i,S\}}H_0(-\frac{\bar{u}-u_k}{b}-z) \end{aligned} \quad (4.8)$$

Using the Taylor expansion on f, we further have

$$h_0(\frac{\bar{u}-u_i}{b}+z)=h_0(z)+h_0'(z)(\frac{\bar{u}-u_i}{b})+\frac{h_0''(z_1)}{2}(\frac{\bar{u}-u_i}{b})^2 \quad (4.9)$$

with $z_1\in(z,\frac{\bar{u}-u_i}{b}+z)$ or $z_1\in(\frac{\bar{u}-u_i}{b}+z,z)$. Similarly, we have

$$H_0(\frac{\bar{u}-u_j}{b}+z)=H_0(z)+h_0(z)(\frac{\bar{u}-u_j}{b})+\frac{h_0'(z_2)}{2}(\frac{\bar{u}-u_j}{b})^2 \quad (4.10)$$

and

$$H_0(-\frac{\bar{u}-u_k}{b}-z)=H_0(-z)-h_0(-z)(\frac{u_k-\bar{u}}{b})-\frac{h_0'(-z_3)}{2}(\frac{u_k-\bar{u}}{b})^2 \quad (4.11)$$

where $z_2\in(z,\frac{\bar{u}-u_j}{b}+z)$ or $z_2\in(\frac{\bar{u}-u_j}{b}+z,z)$, $z_3\in(z,\frac{u_k-\bar{u}}{b}+z)$ or $z_3\in(\frac{u_k-\bar{u}}{b}+z,z)$.

Substituting (4.9), (4.10), and (4.11) into (4.8), following similar argument as one can notice following facts.

1. Summation of all terms multiplied by $1/b$ is

$$\frac{1}{b}\sum_{i=1}^{2n+1}h_0(z)\sum_{S\in\mathcal{S}_i}\prod_{j\in S}H_0(z)\prod_{k\in[n]\setminus\{i,S\}}H_0(-z) \quad (4.12)$$

2. All the terms multiplied by $1/b^2$ cancels with each other after summation due to the definition of $\bar{u}$. I.e.

$$\sum_{i=1}^{2n+1}\frac{1}{b}h_0'(z)(\frac{\bar{u}-u_i}{b})\sum_{S\in\mathcal{S}_i}H_0(z)^nH_0(-z)^n=0 \quad (4.13)$$

$$\sum_{i=1}^{2n+1}\frac{1}{b}h_0(z)\sum_{S\in\mathcal{S}_i}\sum_{j\in S}h_0(z)(\frac{\bar{u}-u_j}{b})H_0(z)^{n-1}H_0(-z)^n=0 \quad (4.14)$$

$$\sum_{i=1}^{2n+1}\frac{1}{b}h_0(z)\sum_{S\in\mathcal{S}_i}H_0(z)^n\sum_{k\in[n]\setminus\{i,S\}}h_0(-z)(\frac{\bar{u}-u_k}{b})H_0(-z)^{n-1}=0 \tag{4.15}$$

3. Excluding the terms above, the rest of the terms are upper bounded by the order of $O(1/b^3)$.

Thus by another change of variable (change z to $\frac{z}{b}$), we have

$$h(\bar{u}+z,b)=\frac{1}{b}g(\frac{z}{b})+\frac{1}{b}v(\frac{z}{b}) \tag{4.16}$$

where

$$g(z)=\sum_{i=1}^{2n+1}h_0(z)\sum_{S\in\mathcal{S}_i}\prod_{j\in S}H_0(z)\prod_{k\in[n]\setminus\{i,S\}}H_0(-z) \tag{4.17}$$

which is the pdf of sample median of $2n+1$ iid draws from h_0 and it is symmetric around 0 .

Further, observe that when $h_0(z)$, $h_0'(z)$, and $h''(z)$ are all absolutely upper bounded and absolutely integrable, integration of absolute value of each high order term in $v(z)$ can be upper bounded in the order of $O(\frac{\max_i|\bar{u}-u_i|^2}{b^2})$. This is because each term in $v(z)$ is at least multiplied by $1/b^2$ and one of $h_0(z)$, $h_0(z)$, $h_0'(z)$, $h_0''(z_1)$, $h_0'(z_2)$ and $h_0'(-z_3)$ (z_1, z_2, z_3 appears through remainder terms of the Taylor' theorem). The terms multiplied by $h_0(z)$, $h_0(-z)$, or $h_0'(z)$ absolutely integrates into a constant. The terms multiplied only by the remainder terms in the integration are more tricky, one need to rewrite the remainder term into integral form and exchange the order of integration to prove that the term integrates the order of $O(1/b^2)$. We do this process for one term in the following and the others are omitted.

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{h_0''(z_1)}{2}(\frac{\bar{u}-u_i}{b})^2H_0(z)H_0(-z)dx\\ \le&\int_{-\infty}^{\infty}\left|\frac{h_0''(z_1)}{2}(\frac{\bar{u}-u_i}{b})^2\right|\|H_0\|_\infty\|H_0\|_\infty dx\\ =&\|H_0\|_\infty\|H_0\|_\infty\int_{-\infty}^{\infty}\left|\int_x^{x+\frac{\bar{u}-u_i}{b}}h_0''(t)(t-x)dt\right|dx\end{aligned} \tag{4.18}$$

where the equality holds because $\frac{h_0''(z_1)}{2}(\frac{\bar{u}-u_i}{b})^2$ is the remainder term of the Taylor expansion when approximating $z+\frac{\bar{u}-u_i}{b}$ at z and we changed the remainder term from the mean-value form to the integral form.

Without loss of generality, we assume $\bar{u} - u_i \geq 0$ (the proof is similar when it is less than 0), then we get

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left| \int_{x}^{x+\frac{\bar{u}-u_i}{b}} h_0''(t)(t-x)dt \right| dx \\
\leq &\int_{-\infty}^{\infty} \int_{x}^{x+\frac{\bar{u}-u_i}{b}} |h_0''(t)||(t-x)|dtdx \\
= &\int_{-\infty}^{\infty} \int_{t-\frac{\bar{u}-u_i}{b}}^{t} \left|h_0''(t)\right||(t-x)|dxdt \\
= &\frac{1}{2}(\frac{\bar{u}-u_i}{b})^2 \int_{-\infty}^{\infty} \left|h_0''(t)\right| dt
\end{aligned} \tag{4.19}$$

which is $(\frac{\bar{u}-u_i}{b})^2$ times a constant.

Thus, we have

$$\int_{-\infty}^{\infty} \frac{1}{b}|v(\frac{z}{b})|dz = \int_{-\infty}^{\infty} \frac{1}{b}|v(z)|bdz = O(\frac{\max_i |\bar{u}-u_i|^2}{b^2}) \tag{4.20}$$

which completes this part. The whole proof finishes here.

Eq. (4.1) is one key result of Theorem 4, i.e., the difference between the expected median and mean shrinks with a rate of $O(1/b)$ as b grows. Another key result of the distribution is (4.5), which says the pdf of the expected median becomes increasingly symmetric and the asymmetric part diminishes with a rate of $O(1/b^2)$. It is worth mentioning that Gaussian distribution satisfies all the assumptions in Theorem 4. In addition, although the theorem is based on assumptions on the second-order differentiability of the pdf, we observe empirically that, many commonly used symmetric distribution with non-differentiable points such as Laplace distribution and uniform distribution can also make the pdf increasingly symmetric and make the expected median closer to the mean, as b increases.

Chapter 5

Convergence of Noisy signSGD and Noisy medianSGD

From Chapter 4, we see that the gap between expected median and mean will be reduced if noise is added. Meanwhile, from the analysis in Chapter 3, MEDIANSGD and SIGNSGD will finally converge to some solution whose gradient size is proportional to the above median-mean gap. Then, a natural idea is to use the noise perturbation mechanism to improve the performance of MEDIANSGD and SIGNSGD.

In this chapter, we propose and analyze the noisy variants of SIGNSGD and MEDIANSGD, where symmetric and unimodal noises are injected on the local gradients. The only difference between the noisy algorithms and the original algorithms is that some artificial noise is added to the stochastic gradients before further processing, i.e. changing line 2 of Algorithm 1 and Algorithm 2 (see Algorithm 3 and Algorithm 4 for pseudo code).

Algorithm 3 Noisy SIGNSGD

1: **Input:** learning rate δ, current point x_t
2: $g_{t,i} = \hat{g}_{t,i} + b\xi_{t,i}$
3: $x_{t+1} \leftarrow x_t - \delta \, \mathrm{sign}(\sum_{i=1}^{M} \mathrm{sign}(g_{t,i}))$

Algorithm 4 Noisy MEDIANSGD

1: **Input:** learning rate δ, current point x_t
2: $g_{t,i} = \hat{g}_{t,i} + b\xi_{t,i}$
3: $x_{t+1} \leftarrow x_t - \delta \mathrm{median}(\{g_{t,i}\}_{i=1}^{M})$

Remark on the sampling noise and connection to differential privacy: The above algorithms still follow the update rule of SIGNSGD and MEDIANSGD, just that the noise on the gradients follows some distribution with good properties described in Chapter 4. Essentially we want the gradient noise to be symmetric. If the noise generated by data sub-sampling is approximate symmetric, this sub-sampling noise should also help with the convergence performance. It is shown in Bernstein et al. [2018a] that the gradient noise generated by data sub-sampling indeed follows some symmetric structure and we show later in Chapter 6 that sub-sampling indeed helps in practice under the situation of heterogeneous data. One may notice the noise $\xi_{t,i}$ is similar to the noise added to ensure differential privacy when it is Gaussian. If an upper bound on gradient norm is known, one can use standard privacy accountant to calculate the privacy cost. Our convergence result implies that certain amount of noise added for differential privacy can help instead of hurt. However, the theoretical utilities of these algorithms under the privacy setting are unknown and we leave them for future works.

Now we present the convergence results for Algorithm 3 and Algorithm 4.

Theorem 5. *Suppose A1, A3, A4 are satisfied and* $|(\hat{g}_{t,i})_j| \leq Q, \forall t, j$. *When each* $\xi_{t,i}$ *is sampled iid from a symmetric and unimodal distribution with mean 0, variance 1 and pdf* $h(z)$. *If* $h'(z)$ *and* $h''(z)$ *are uniformly bounded and absolutely integrable. Define* σ *to be standard deviation of median of* $2n+1$ *samples drawn from* $h(z)$, $D_f \triangleq f(x_1) - \min_x f(x)$, $\mathcal{W}_t$ *to be the set of coordinates* j *at iteration* t *with* $\frac{|\nabla f(x_t)_j|}{b\sigma} \geq \frac{2}{\sqrt{3}}$. *For* SIGNSGD, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j\in\mathcal{W}_t}|\nabla f(x_t)_j| + \sum_{j\in[d]\setminus\mathcal{W}_t}\frac{1}{b\sigma}\nabla f(x_t)_j^2\right)$$
$$\leq \frac{3D_f}{T\delta} + \frac{3}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{d}|\nabla f(x_t)_j| O\left(\frac{1}{b^2}\right)\right] + \frac{3L}{2}\delta d \tag{5.1}$$

Effect of b and comparison with error feedback. Theorem 5 shows how different parameters can affect the convergence guarantee. The term that decreases with b on RHS of (5.1) is the convergence gap introduced by heterogeneous data and data sampling. Without artificial noise, this term is a constant. Such a term implies adding noise can provably improve the performance of SIGNSGD if it is stuck due to the median-mean gap. In the proof, this gap appears when bounding asymetricity of distribution of the median gradient, which can decrease after adding noise and can be small on a homogeneous distribution (see (5.15)). The other two terms on the RHS of (5.1) are standard in convergence analysis and can diminish with proper T and δ. The trade-off between convergence speed and final accuracy is reflected on LHS of (5.1), i.e. the algorithm can converge better but slower with larger noise. If one optimizes b and δ w.r.t T and d, one can get a worst case bound $\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j\in\mathcal{W}_t} T^{1/4}d^{1/4}|\nabla f(x_t)_j| + \sum_{j\in[d]\setminus\mathcal{W}_t}\nabla f(x_t)_j^2\right) \leq O\left(\frac{d^{3/4}}{T^{1/4}}\right)$ by setting $b = T^{1/4}d^{1/4}$ and $\delta = 1/\sqrt{Td}$. Such a rate is slower than SGD's $O(\sqrt{d}/\sqrt{T})$ rate, and it implies that signSGD may not be preferred when data heterogeneity is large. Compared with the error feedback fix of signSGD in Karimireddy et al. [2019]. Our approach address the convergence issue in multi-worker setting and does not modify update rule signSGD, while the error feedback fix is designed for single-worker setting and uses gradient magnitude information in parameter updates.

Proof: Use the fact that the noise on median is approximately unimodal and symmetric, one may prove that signSGD can converge to a stationary point. With symmetric and unimodal noise, the bias in signSGD can be alternatively viewed as a decrease of effective learning rate, thus slowing down the optimization instead of leading a constant bias. This proof formalizes this idea by characterizing the asymmetricity of the noise ($O(1/\sigma^2)$) and then follows a sharp analysis for signSGD. The key difference from Theorem 1 is taking care of the bias introduced by the difference between median and mean.

Let us recall:

$$\text{median}(\{g_t\}) \triangleq \text{median}(\{(g_{t,i}\}_{i=1}^{M}), \tag{5.2}$$

$$\text{median}(\{\nabla f_t\}) \triangleq \text{median}(\{\nabla f_i(x_t)\}_{i=1}^{M}). \tag{5.3}$$

where

$$g_{t,i} = \nabla f_i(x_t) + b\xi_{t,i} + \zeta_{t,i} \tag{5.4}$$

where $\xi_{t,i}$ is a d dimensional random vector with each element drawn iid from $N(0,1)$ and we abuse the notation $\zeta_{t,i}$ to denote an zero-mean additive discrete noise caused by sampling on data.

By (3.13), we have the following series of inequalities

$$
\begin{aligned}
& f(x_{t+1}) - f(x_t) \\
\leq & -\delta \langle \nabla f(x_t), \text{sign}(\text{median}(\{g_t\})) \rangle + \frac{L}{2}\delta^2 d \\
= & -\delta \sum_{j=1}^{d} |\nabla f(x_t)_j| (I[\text{sign}(\text{median}(\{g_t\})_j) = \text{sign}(\nabla f(x_t)_j)]) \\
& + \delta \sum_{j=1}^{d} |\nabla f(x_t)_j| (I[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\nabla f(x_t)_j)]) \\
& + \frac{L}{2}\delta^2 d
\end{aligned} \tag{5.5}
$$

where $\text{median}(\{g_t\})_j$ is jth coodrinate of $\text{median}(\{g_t\})$, and $I[\cdot]$ denotes the indicator function.

Taking expectation over all the randomness, we get

$$
\begin{aligned}
& \mathbb{E}[f(x_{t+1})] - \mathbb{E}[f(x_t)] \\
\leq & -\delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| \left(P[\text{sign}(\text{median}(\{g_t\})_j) = \text{sign}(\nabla f(x_t)_j)] \right) \right] \\
& + \delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| \left(P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\nabla f(x_t)_j)] \right) \right] \\
& + \frac{L}{2}\delta^2 d
\end{aligned} \tag{5.6}
$$

Now we need a refined analysis on the error probability. In specific, we need an sharp analysis on the following quantity

$$
P[\text{sign}(\text{median}(\{g_t\})_j) = \text{sign}(\nabla f(x_t)_j)] - P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\nabla f(x_t)_j)]. \tag{5.7}
$$

Using reparameterization, we can rewrite $\text{median}(\{g_t\})$ as

$$
\text{median}(\{g_t\}) = \nabla f(x_t) + \xi_t \tag{5.8}
$$

where ξ_t is created by $\xi_{t,i}$'s and $\zeta_{t,i}$'s added on the local gradients on different nodes.

Then, w.l.o.g., assume $\nabla f(x_t)_j \geq 0$ we have

$$\begin{aligned} & P[\text{sign}(\text{median}(\{g_t\})_j) \neq \nabla f(x_t)_j] \\ =& P[(\xi_t)_j \leq -\nabla f(x_t)_j] \\ =& \int_{-\infty}^{-\nabla f(x_t)_j} h_{t,j}(z) \end{aligned} \tag{5.9}$$

where $h_{t,j}(z)$ is the pdf of the jth coordinate of ξ_t .

Similarly, we have

$$\begin{aligned} & P[\text{sign}(\text{median}(\{g_t\})_j) = \nabla f(x_t)_j] \\ =& P[(\xi_t)_j > -\nabla f(x_t)_j] \\ =& \int_{-\nabla f(x_t)_j}^{\infty} h_{t,j}(z) \end{aligned} \tag{5.10}$$

From (4.4) and (4.5), we can split $h_{t,j}(z)$ into a symmetric part and a non-symmetric part which can be written as

$$h_{t,j}(z) = h_{t,j}^s(z) + h_{t,j}^u(z) \tag{5.11}$$

where $h_{t,j}^s(z)$ is symmetric around 0 and $h_{t,j}^u(z)$ is not.

Therefore, from (5.10) and (5.9), we know that

$$\begin{aligned} & P[\text{sign}(\text{median}(\{g_t\})_j) = \text{sign}(\nabla f(x_t)_j)] - P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\nabla f(x_t)_j)] \\ =& \int_{-\nabla f(x_t)_j}^{\infty} h_{t,j}(z) - \int_{-\infty}^{-\nabla f(x_t)_j} h_{t,j}(z) \\ =& \int_{-\nabla f(x_t)_j}^{\infty} h_{t,j}^s(z) + h_{t,j}^u(z) - \int_{-\infty}^{-\nabla f(x_t)_j} h_{t,j}^s(z) + h_{t,j}^u(z) \\ =& \int_{-\nabla f(x_t)_j}^{\nabla f(x_t)_j} h_{t,j}^s(z) + \int_{-\nabla f(x_t)_j}^{\infty} h_{t,j}^u(z) - \int_{-\infty}^{-\nabla f(x_t)_j} h_{t,j}^u(z) \end{aligned} \tag{5.12}$$

where the last equality is due to symmetricity of $h_{t,j}^s(z)$, and the assumption that $\nabla f(x_t)_j$ is positive.

To simplify the notations, define a new variable $z_{t,j}$ with pdf $h_{t,j}^s$, then we have

$$\int_{-\nabla f(x_t)_j}^{\nabla f(x_t)_j} h_{t,j}^s(z) = P[|z_{t,j}| \leq |\nabla f(x_t)_j|] \tag{5.13}$$

A similar result can be derived for $\nabla f(x_t)_j \leq 0$.

In addition, since the noise on each coordinate of local gradient satisfy Theorem 4, we can apply (4.5) to each coordinate of the stochastic gradient vectors. Denote $\bar{\hat{g}}_t = \frac{1}{2n+1}\sum_{i=1}^{2n+1} \hat{g}_{t,i}$ we know that

$$\int_{-\infty}^{\infty} |h^u_{t,j}(z)| = O\left(\frac{\mathbb{E}_{\{\hat{g}_{t,i}\}_{i=1}^{2n+1}}[\max_i |(\hat{g}_{t,i})_j - (\bar{\hat{g}}_t)_j|^2]}{b^2}\right) + O\left(\frac{\mathbb{E}_{\{\hat{g}_{t,i}\}_{i=1}^{2n+1}}((\bar{\hat{g}}_t)_j - \nabla f(x_t)_j)^2}{b^2}\right) \tag{5.14}$$

and thus

$$\begin{aligned} &\int_{-\nabla f(x_t)_j}^{\infty} h^u_{t,j}(z) - \int_{-\infty}^{-\nabla f(x_t)_j} h^u_{t,j}(z) \\ \leq &O\left(\frac{\mathbb{E}_{\{\hat{g}_{t,i}\}_{i=1}^{2n+1}}[\max_i |(\hat{g}_{t,i})_j - (\bar{\hat{g}}_t)_j|^2]}{b^2}\right) + O\left(\frac{\mathbb{E}_{\{\hat{g}_{t,i}\}_{i=1}^{2n+1}}((\bar{\hat{g}}_t)_j - \nabla f(x_t)_j)^2}{b^2}\right) \\ = &O(\frac{1}{b^2}). \end{aligned} \tag{5.15}$$

We comment that (5.15) can be small if a large minibatch is used and $\max_i \|\nabla f_i(x_t) - \nabla f(x_t)\|$ is small even with a constant b. Consider when full batch gradient evaluation is used. In this case, we know $\hat{g}_{t,i} = \nabla f_i(x_t)$ and the second term on RHS of the first inequality of (5.15) become 0. The first term becomes $O\left(\frac{\max_i |\nabla f_i(x_t)_j - \nabla f(x_t)_j|^2}{b^2}\right)$ which can be bounded by $\max_i \|\nabla f_i(x_t) - \nabla f(x_t)\|$. With homogeneous data distribution one usually have $\max_i \|\nabla f_i(x_t) - \nabla f(x_t)\|$ being a small number. Getting back to the minibatch case where the stochastic gradients are evaluated on a minibatches, when the number of samples in a batch is large, we have $\hat{g}_t$ being close to its mean $\nabla f(x_t)$ with high probability by the law of large numbers which means the second term on RHS of (5.15) is small. Similarly, we have $\hat{g}_{t,i}$ approaching $\nabla f_i(x_t)$ and the first term becomes similar to the full batch case, which be small on a heterogeneous data distribution.

To continue, we need to introduce some new definitions. Define

$$W_{t,j} = \frac{|\nabla f(x_t)_j|}{b\sigma_{mid}} \tag{5.16}$$

where σ_{mid} is the variance of the noise with pdf (4.4), that is

$$g(z) = \sum_{i=1}^{2n+1} h_0(z) \sum_{S\in\mathcal{S}_i} \prod_{j\in S} H_0(z) \prod_{k\in[n]\setminus\{i,S\}} H_0(-z) \tag{5.17}$$

By adapting Lemma 1 from Bernstein et al. [2018b] (which is an application of Gauss's inequality), we have

$$P[|z_{t,j}| < |\nabla f(x_t)_j|] \geq \begin{cases} 1/3 & W_{t,j} \geq \frac{2}{\sqrt{3}} \\ \frac{W_{t,j}}{\sqrt{3}} & \text{otherwise} \end{cases} \tag{5.18}$$

Thus, continuing from (5.6), we have

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - \mathbb{E}[f(x_t)] \\
\leq &-\delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| (P[\text{sign}(\text{median}(\{g_t\})_j) = \text{sign}(\nabla f(x_t)_j)])\right] \\
&+ \delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| (P[\text{sign}(\text{median}(\{g_t\})_j) \neq \text{sign}(\nabla f(x_t)_j)])\right] \\
&+ \frac{L}{2}\delta^2 d \\
\leq &-\delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| (P[|z_{t,j}| < |\nabla f(x_t)_j|])\right] \\
&+ \delta \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| O\left(\frac{1}{b^2}\right)\right] + \frac{L}{2}\delta^2 d
\end{aligned} \tag{5.19}$$

Define $D_f \triangleq f(x_1) - \min_x f(x)$, telescope from 1 to T, divide both sides by $T\delta$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| (P[|z_{t,j}| < |\nabla f(x_t)_j|])\right] \\
&\leq \frac{1}{T\delta} D_f + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| O\left(\frac{1}{b^2}\right)\right] + \frac{L}{2}\delta d
\end{aligned} \tag{5.20}$$

where the RHS is decaying with a speed of $\frac{\sqrt{d}}{\sqrt{T}}$.

Further, substituting (5.18) and (5.16) into (5.20) and multiplying both sides of (5.20) by 3, we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j \in \mathcal{W}_t} |\nabla f(x_t)_j| + \frac{1}{b\sigma_{mid}} \sum_{j \in [d] \setminus \mathcal{W}_t} \nabla f(x_t)_j^2\right) \\
&\leq 3\frac{1}{T\delta} D_f + 3\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{d} |\nabla f(x_t)_j| O\left(\frac{1}{b^2}\right)\right] + 3\frac{L}{2}\delta d
\end{aligned} \tag{5.21}$$

which completes the proof.

Theorem 6. *Suppose A1, A3, A4 are satisfied and* $|(\hat{g}_{t,i})_j| \leq Q, \forall t, j$. *When each coordinate of* $\xi_{t,i}$ *is sampled iid from a symmetric and unimodal distribution with mean 0, variance 1 and pdf* $h(z)$. *If* $h'(z)$ *is uniformly bounded and absolutely integrable. Define* $D_f \triangleq f(x_1) - \min_x f(x)$, *set* $\delta \leq \frac{1}{2L}$. *For* MEDIANSGD, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2}{T\delta}D_f + O\left(\frac{d}{b^2}\right) + O\left(\delta d b^2\right) \tag{5.22}$$

Effect of b and robustness after perturbation. The trade-off between convergence speed and convergence accuracy in Theorem 6 is more clear compared with Theorem 5. Larger b can effectively reduce the median-mean gap and simultaneously induce a larger noise on gradient. The best convergence rate is $O(d^{2/3}/T^{1/3})$ by setting $b = T^{1/6}d^{1/6}$ and $\delta = 1/(T^{2/3}d^{2/3})$. The robustness of MEDIANSGD comes from the fact that when performing variable updates, the mean of gradients is replaced by the median of gradients, which is less sensitive to extreme values. Since the noise perturbation technique is gradually converting the median estimator to a mean estimator as more noise is added, it should make MEDIANSGD less robust simultaneously. On heterogeneous data with Byzantine workers, the performance of the median-based algorithms can be affected by mainly two factors. One is the gap between median and mean of gradients, the other one is the misleading information provided by Byzantine workers. When the possible effect of Byzantine workers is small (e.g. a small number of Byzantine workers) compared with the median-mean gap, some noise might be still preferred to reduce the median-mean gap even though this could amplify the effect of Byzantine workers. For non-heterogeneous data (e.g. the data is distributed iid across workers), the median of gradient could be very close to the mean especially when the number of training samples is large. In such cases, the median-mean gap may not be a bottleneck for the performance of the algorithm and the noise standard deviation b should be set small or even be 0. There could be a trade-off between convergence guarantee and robustness and a thorough quantitative study is left for future works.

Proof: Following the same procedures as the Theorem 2, we can get

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - \mathbb{E}[f(x_t)] \\
\leq &-(\frac{\delta}{2} - L\delta^2)\mathbb{E}[\|\mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] + \frac{\delta}{2}\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] \\
&+ L\delta^2\mathbb{E}[\|\text{median}(\{g_t\}) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] - \frac{\delta}{2}\mathbb{E}[\|\nabla f(x_t)\|^2]
\end{aligned} \tag{5.23}$$

which is the same as (3.23).

Sum over $t \in [T]$ and divide both sides by $T\delta/2$, assume $\delta \leq \frac{1}{2L}$, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \\
\leq &\frac{2}{T\delta}(\mathbb{E}[f(x_1)] - \mathbb{E}[f(x_{T+1})]) + +\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\|^2] + 2Ld\delta\sigma_m^2
\end{aligned} \tag{5.24}$$

By (A.1) in Theorem A.1 (which is a variant of Theorem 4 with minibatch sampling), we know

$$\mathbb{E}[\|\nabla f(x_t) - \mathbb{E}[\text{median}(\{g_t\})|x_t]\| = O(\frac{\sqrt{d}}{b}) \tag{5.25}$$

where $\sqrt{d}$ is due to L_2 norm. In addition, we have $\sigma_m^2 = O(b^2)$ by (A.2) in Theorem A.1. Assume $\mathbb{E}[\text{median}(\{g_t\})_j|x_t] \leq Q$ and set $b = T^{1/6}d^{1/6}$ and $\delta = T^{-2/3}d^{-2/3}$, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \\
\leq &\frac{2}{T\delta}(\mathbb{E}[f(x_1)] - \mathbb{E}[f(x_{T+1})]) + O\left(\frac{d}{b^2}\right) + O\left(\delta d b^2\right).
\end{aligned} \tag{5.26}$$

Then, upper bounding $\mathbb{E}[f(x_1)] - \mathbb{E}[f(x_{T+1})])$ by D_f finishes the proof.

Chapter 6

Experiments

In this chapter, we show how adding noise helps the practical behavior of the algorithms. Since signSGD is better studied empirically and medianSGD is more of theoretical interest so far, we use signSGD to demonstrate the benefit of injecting noise. We conduct experiments on MNIST and CIFAR-10 datasets. For both datasets, the data distribution on each node is heterogeneous, more specifically, each node contains some exclusive data for one or two out of ten categories.

Our experimentation is mainly implemented using Python 3.6.4 with packages including MPI4Py 3.0.0, NumPy 1.14.2 and TensorFlow 1.10.0. We use the Message Passing Interface (MPI) to implement the distributed system, and use TensorFlow to implement the neural network. The MNIST experiments are run on up to 20 compute cores of two Intel Haswell E5-2680 CPUs with 64 GB Memory. The CIFAR-10 experiments are run on up to 5 AWS p3.2xlarge machines.

Dataset and pre-processing: In the first experiment, we use the MNIST dataset[1] , which contains a training set of 60,000 samples, and a test set of 10,000 samples, both are 28x28 grayscale images of the 10 handwritten digits. To facilitate the neural network training, the original feature vector, which contains the integer pixel value from 0 to 255, has been scaled to a float vector in the range $(0, 1)$. The integer categorical label is also converted to the binary class matrix (one hot encoding) for use with the categorical cross-entropy loss. For the CIFAR-10 dataset, the data are processed in the same way.

Neural Network and Initialization: For MNIST, a two-layer fully connected

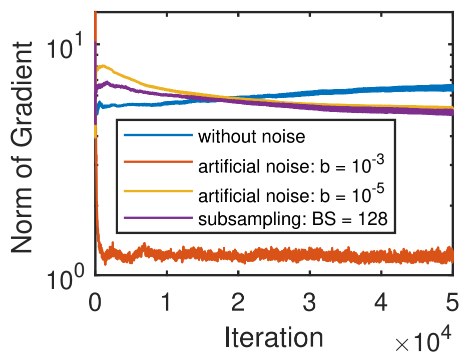
[1] Available at http://yann.lecun.com/exdb/mnist/

neural network with 128 and 10 neurons for each layer is used in the experiment. The initialization parameters are drawn from a truncated normal distribution centered on zero, with variance scaled with the number of input units in the weight tensor (fan-in). For CIFAR-10, we use ResNet-20 (obtained from the implementation of ResNet20 v1 in Keras) with batch normalization layers removed, parameters are initialized the same way as in Keras. We removed the batch normalization layers because inconsistency of statistics (mean and variance) of batch normalization layers on different nodes significantly deteriorates the performance when heterogeneous data is used. This phenomenon is also observed and explained in Hsieh et al. [2019], Ioffe [2017].

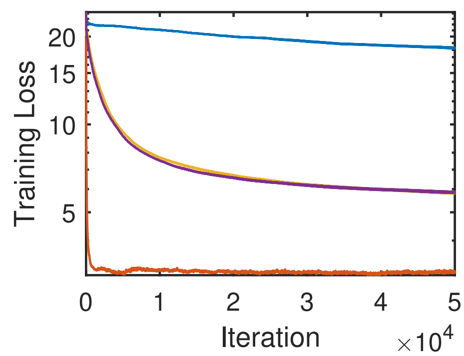
Parameter Tuning: We use constant stepsize for MNIST and a learning rate schedule for CIFAR-10. For MNIST, the stepsize is chosen from the set {1, 0.1, 0.01, 0.001} based on training performance. For CIFAR-10, the initial learning rate is chosen from the set {1, 0.1, 0.01, 0.001}, the stepsize is divided by 2, 10, 20 after 1000, 3000, 5000 iterations, respectively. For CIFAR-10, the standard deviation of the added noise (b in the algorithm) is set to be different for different weights. Specifically, for weight W_i, we set $b = 0.1 * \max_j(Q_{i,j})$ where $Q_{i,j}$ is the maximum of absolute value of elements in stochastic gradient w.r.t W_i at node j. This requires transmitting an additional float number per weight but the cost is negligible given the sizes of the weights. We found the aforementioned adaptive noise adding scheme makes it easier to tune the noise level.

We first compare full batch Noisy signSGD (Algorithm 3) with different b (i.e. adding random Gaussian noise with different standard deviation), signSGD with sub-sampling on data , and full batch signSGD without any noise on MNIST. This experiment is to check the effects of the noise generated from data sub-sampling and the artificial noise. We see that signSGD without noise stuck at some point where the gradient is a constant (the median should be oscillating around zero). At the same time, both Noisy signSGD or signSGD with sub-sampling drive the gradient to be smaller. From the results in Figure 6.1, we can see that both training accuracy and test accuracy of signSGD without noise are very poor. The perturbation indeed helps improve the classification accuracy in practice.

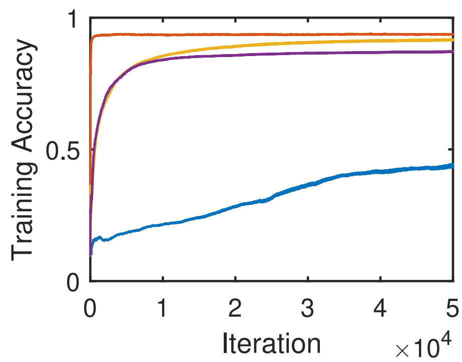
In the second experiment (see Figure 6.2), we examine the performance of Noisy signSGD on CIFAR-10. We compare signSGD on both heterogeneous and homogeneous data, and Noisy signSGD on heterogeneous data. We can see that when artificial



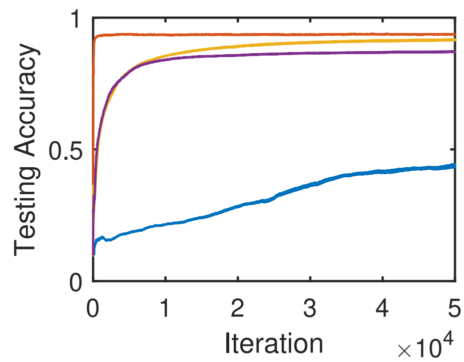
(a) Norm of Gradient



(b) Training Loss

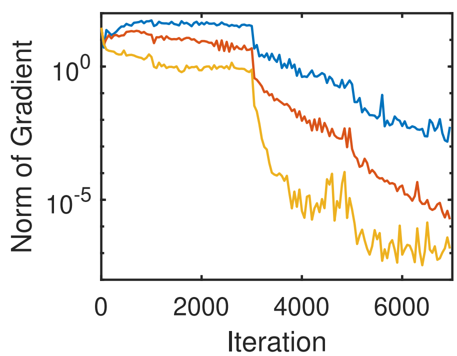


(c) Training Accuracy

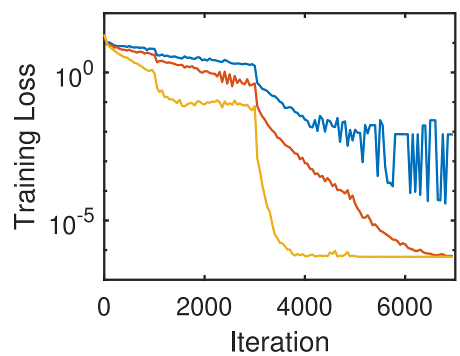


(d) Testing Accuracy

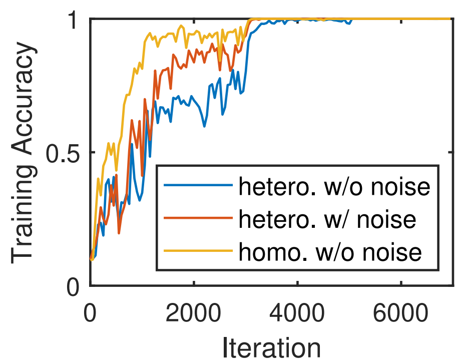
Figure 6.1: Comparison of SIGNSGD with different noise on MNIST.



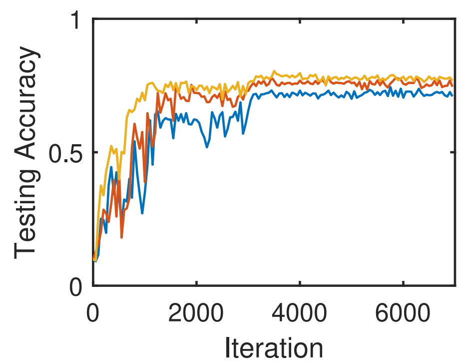
(a) Norm of Gradient



(b) Training Loss



(c) Training Accuracy



(d) Testing Accuracy

Figure 6.2: Comparison of SIGNSGD on CIFAR-10. For the noisy algorithms we use $b = 0.001$. The sudden change of performance is caused by learning rate decay, which happens at 1000/3000/5000 iterations.

noise is added to the gradients, the convergence speed of signSGD is significantly increased. In addition, we can see that a side benefit of adding noise is improving generalization performance, despite all the algorithm achieves almost 100% training accuracy, Noisy signSGD achieves higher testing accuracy than vanilla signSGD. It is shown that signSGD on homogeneous data converges with similar speed of SGD in Bernstein et al. [2018a] while reducing 16x or 32x communication per iteration. Comparing convergence speed of Noisy signSGD on heterogeneous data and signSGD on homogeneous data, one can argue Noisy signSGD should also be communication efficient when considering total communication compared with SGD. This set of experiments again shows that injecting noise does help in practice.

Chapter 7

Conclusion and Discussion

In this paper, we uncover the connection between signSGD and medianSGD by showing signSGD is a median-based algorithm. We also show that when the data at different nodes come from different distributions, the class of median-based algorithms suffers from non-convergence caused by using the median to evaluate mean. To fix the non-convergence issue, We provide a perturbation mechanism to shrink the gap between the expected median and mean. By incorporating the perturbation mechanism, we show the convergence of both signSGD and medianSGD is guaranteed to improve. To the best of our knowledge, this is the first time that median-based methods, including signSGD and medianSGD, are able to converge with a provable rate for distributed problems with heterogeneous data. The perturbation mechanism can be approximately realized by sub-sampling of data during gradient evaluation, which partly supports the use of sub-sampling in practice. We also conducted experiments on training neural nets to show the necessity of the perturbation mechanism and sub-sampling.

References

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*, pages 1709–1720, 2017.

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4613–4623, 2018.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. *arXiv preprint arXiv:1807.00459*, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 559–568, 2018a.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018b.

Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 119–129, 2017.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip B Gibbons. The non-iid data quagmire of decentralized machine learning. *arXiv preprint arXiv:1910.00189*, 2019.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in neural information processing systems*, pages 1945–1953, 2017.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation*, OSDI'14, pages 583–598, Berkeley, CA, USA, 2014. USENIX Association. ISBN 978-1-931971-16-4.

Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.

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 International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

Steven J Miller. *The Probability Lifesaver: Order Statistics and the Median Theorem.* Princeton University Press, 2017.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. *arXiv preprint arXiv:1909.13014*, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 2019.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309, 2018.

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. In *Advances in neural information processing systems*, pages 1509–1519, 2017.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5636–5645, 2018.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.

Appendix A

An extended version of Theorem 4

In the proof of convergence of SIGNSGD and MEDIANSGD, we need a version of Theorem 4 with minibatch sampling of training samples at each node, we present the theorem and its proof in this chapter.

Theorem A.1. *Assume we have 2n+1 set of numbers with each set denoted by* $A_i = \{a_{i,j}\}_{j=1}^{k_i}, i \in [2n+1]$ *with mean of the numbers of each set being* $u_1, .., u_{2n+1}$. *Given a symmetric and unimodal noise distribution with mean 0, variance 1. Denote the pdf of the distribution to be* $h_0(z)$ *and cdf to be* $H_0(z)$. *Suppose* $h_0'(z)$ *is uniformly bounded and absolutely integrable. Draw* $2n+1$ *samples* $\xi_1, ..., \xi_{2n+1}$ *from the distribution* $h_0(z)$. *Define random variable* q_i *to be a number uniformly randomly drawn from* $\{a_{i,j}\}_{j=1}^{K_i}$ *and* $\hat{q}_i = q_i + b\xi_i$, $\bar{u} \triangleq \sum_{i=1}^{2n+1} u_i$,

(a) We have

$$\mathbb{E}[\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})] = \bar{u} + O\left(\frac{\max_{i,j,i',j'|i\neq i'} |a_{i,j} - a_{i',j'}|^2}{b}\right), \tag{A.1}$$

$$\text{Var}(\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})) = O(b^2). \tag{A.2}$$

(b) Further assume $h_0''(z)$ *is uniformly bounded and absolutely integrable. Denote* $r_b(z)$ *to be the pdf of the distribution of* $\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})$ *and* $S_A = A_1 \times A_2 \times ... \times A_{2n+1}$,

we have

$$r_b(\bar{u}+z) = \underbrace{\frac{1}{b}g(\frac{z}{b})}_{\text{symmetric part}} + \underbrace{\frac{1}{b}v(\frac{z}{b})}_{\text{asymmetric part}} \tag{A.3}$$

where

$$g(z) \triangleq \sum_{i=1}^{2n+1} h_0(z) \sum_{S\in\mathcal{S}_i} \prod_{j\in S} H_0(z) \prod_{k\in[n]\setminus\{i,S\}} H_0(-z) \tag{A.4}$$

being the pdf of sample median of $2n+1$ samples drawn from the distribution $h_0(z)$ which is symmetric over 0, $\mathcal{S}_i$ is the set of all n-combinations of items from the set $[2n+1]\setminus i$, and the asymmetric part satisfies

$$\int_{-\infty}^{\infty} \frac{1}{b}|v(\frac{z}{b})|dz = O\left(\frac{\mathbb{E}_{s\sim U(S_A)}[\max_i |s_i-\bar{s}|^2]}{b^2}\right) + O\left(\frac{\max_{s\in S_A}(\bar{s}-\bar{u})^2}{b^2}\right) \tag{A.5}$$

where $U(S_A)$ is uniform distribution over elements in S_A

Proof of Theorem A.1: The proof is mostly based on Theorem 4 with some extra efforts dealing with the sampling noise. We first prove part (a) (A.1). Since q_i is sampled uniformly randomly from $\{a_{i,j}\}_{j=1}^{K_i}$, we know there are $\prod_{i=1}^{2n+1} K_i$ possible realizations for $\{q_i\}_{i=1}^{2n+1}$ with equal probability. I addition, the mean of mean of these realizations is $\bar{u}$. For each realization $\{q_i = \tilde{q}_i\}_{i=1}^{2n+1}$, we know from Theorem 4 (a) that

$$\mathbb{E}[\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i=\tilde{q}_i\}_{i=1}^{2n+1}] = \frac{1}{2n+1}\sum_{i=1}^{2n+1}\tilde{q}_i + O\left(\frac{\max_{i,i'}|\tilde{q}_i-\tilde{q}_{i'}|^2}{b}\right). \tag{A.6}$$

Given the fact that

$$\mathbb{E}[\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})] = \mathbb{E}_{\{q_i\}_{i=1}^{2n+1}}[\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i\}_{i=1}^{2n+1}]]$$

and $\mathbb{E}[\frac{1}{2n+1}\sum_{i=1}^{2n+1} q_i] = \bar{u}$, we know

$$\mathbb{E}[\text{median}(\{\hat{q}_i\}_{i=1}^{2n+1})] = \bar{u} + O\left(\frac{\max_{i,i',j,j',i\neq i'}|a_{i,i}-a_{i',j'}|^2}{b}\right). \tag{A.7}$$

which proves (A.1).

Now we prove (A.2). We have

$$
\begin{aligned}
&\mathrm{Var}(\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1}))\\
=&\mathbb{E}[(\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1}) - \mathbb{E}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})])^2]\\
=&\mathbb{E}_{\{q_i\}_{j=1}^{2n+1}}[\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[(\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1}) - \mathbb{E}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})])^2|\{q_i\}_{j=1}^{2n+1}]]\\
\leq&\mathbb{E}_{\{q_i\}_{j=1}^{2n+1}}[\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[2(\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})\\
&-\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i\}_{j=1}^{2n+1}])^2|\{q_i\}_{j=1}^{2n+1}]]\\
&+\mathbb{E}_{\{q_i\}_{j=1}^{2n+1}}[\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[2(\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i\}_{j=1}^{2n+1}]\\
&-\mathbb{E}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1}])^2|\{q_i\}_{j=1}^{2n+1}]]\\
=&2\mathbb{E}_{\{q_i\}_{j=1}^{2n+1}}[\mathrm{Var}(\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i\}_{j=1}^{2n+1})]\\
&+2\mathbb{E}_{\{q_i\}_{j=1}^{2n+1}}[(\mathbb{E}_{\{\xi_i\}_{j=1}^{2n+1}}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1})|\{q_i\}_{j=1}^{2n+1}] - \mathbb{E}[\mathrm{median}(\{\hat{q}_i\}_{i=1}^{2n+1}])^2]\\
\leq&O(b^2)+O(1)
\end{aligned}
\tag{A.8}
$$

where the last inequality is because we can apply (4.2) to the variance term and for the remaining term we have (A.6) and (A.7).

Now we prove part (b) of the theorem. Denote $A_i = \{a_i\}_{i=1}^{K_i}$ and $S_A = A_1 \times A_2 \times ... \times A_{2n+1}$. Also, given a vector $s \in R^{2n+1}$, denote $r_b(z, s)$ to be the pdf of median$(\{\hat{q}_i\}_{i=1}^{2n+1})$ with $q_i = s_i, \forall i \in [2n+1]$. By definition of median$(\{\hat{q}_i\}_{i=1}^{2n+1})$, we know know

$$
r_b(z) = \frac{1}{|S_A|}\sum_{s\in S_A} r_b(z,s). \tag{A.9}
$$

Denote $\bar{s} = \frac{1}{2n+1}\sum_{i=1}^{2n+1} s_i$, from Theorem 4 (b), we know

$$
r_b(\bar{s}+z,s) = \underbrace{\frac{1}{b}g(\frac{z}{b})}_{\text{symmetric part}} + \underbrace{\frac{1}{b}v(\frac{z}{b})}_{\text{asymmetric part}}
$$

with

$$
g(z) = \sum_{i=1}^{2n+1} h_0(z) \sum_{S\in\mathcal{S}_i}\prod_{j\in S} H_0(z) \prod_{k\in[n]\setminus\{i,S\}} H_0(-z)
$$

being the pdf of sample median of $2n+1$ samples drawn from the distribution $h_0(z)$ which is symmetric over 0, $\mathcal{S}_i$ is the set of all n-combinations of items from the set

$[2n+1] \setminus i$, and

$$\int_{-\infty}^{\infty} \frac{1}{b}|v(\frac{z}{b})|dz = O\left(\frac{\max_i |\bar{s}-s_i|^2}{b^2}\right)$$

By Taylor expansion, we know

$$\begin{aligned} r_b(\bar{u}+z, s) =& \frac{1}{b}g(\frac{z+\bar{s}-\bar{u}}{b}) + \frac{1}{b}v(\frac{z+\bar{s}-\bar{u}}{b}) \\ =& \frac{1}{b}\left(g(\frac{z}{b}) + g'(\frac{z}{b})(\frac{\bar{s}-\bar{u}}{b}) + \frac{1}{2}g''(\frac{z_1}{b})(\frac{\bar{s}-\bar{u}}{b})^2\right) + \frac{1}{b}v(\frac{z+\bar{s}-\bar{u}}{b}) \end{aligned} \tag{A.10}$$

with $z_1 \in [\min(z, z+\bar{s}-\bar{u}), \max(z, z+\bar{s}-\bar{u})]$

Substituting (A.10) into (A.9), we know

$$\begin{aligned} & r_b(\bar{u}+z) \\ =& \frac{1}{|S_A|}\sum_{s\in S_A} r_b(\bar{u}+z, s) \\ =& \frac{1}{|S_A|}\sum_{s\in S_A}\frac{1}{b}\left(g(\frac{z}{b}) + g'(\frac{z}{b})(\frac{\bar{s}-\bar{u}}{b}) + \frac{1}{2}g''(\frac{z_1}{b})(\frac{\bar{s}-\bar{u}}{b})^2\right) + \frac{1}{|S_A|}\sum_{s\in S_A}\frac{1}{b}v(\frac{z+\bar{s}-\bar{u}}{b}) \\ =& \frac{1}{b}g(\frac{z}{b}) + \frac{1}{|S_A|}\sum_{s\in S_A}\frac{1}{b}\left(\frac{1}{2}g''(\frac{z_1}{b})(\frac{\bar{s}-\bar{u}}{b})^2\right) + \frac{1}{|S_A|}\sum_{s\in S_A}\frac{1}{b}v(\frac{z+\bar{s}-\bar{u}}{b}) \end{aligned} \tag{A.11}$$

where we have used the fact that $\frac{1}{|S_A|}\sum_{s\in S_A}\bar{s} = \bar{u}$ in the last equality.

What remains is to bound the integration of terms other than $\frac{1}{b}g(\frac{z}{b})$ in (A.11).

From Theorem 4 (b), we already know

$$\begin{aligned} & \int_{-\infty}^{\infty}\frac{1}{|S_A|}\sum_{s\in S_A}\left|\frac{1}{b}v(\frac{z+\bar{s}-\bar{u}}{b})\right| dz \\ \leq& \frac{1}{|S_A|}\sum_{s\in S_A} O\left(\frac{\max_i |s_i-\bar{s}|^2}{b^2}\right) \\ =& O\left(\frac{1}{|S_A|}\sum_{s\in S_A}\frac{\max_i |s_i-\bar{s}|^2}{b^2}\right) \end{aligned} \tag{A.12}$$

Thus, we only need to bound $\frac{1}{|S_A|}\frac{1}{2b}\sum_{s\in S_A}\int_{-\infty}^{\infty}\left|g''(\frac{z_1}{b})(\frac{\bar{s}-\bar{u}}{b})^2\right| dz$. Each term in the summation can be upper bounded as $(\frac{(\bar{s}-\bar{u})^2}{b^2}\int_{-\infty}^{\infty}|g''(z)|d_z$ using the same procedure in

(4.18) and (4.19). Thus, we have

$$\frac{1}{|S_A|}\sum_{s\in S_A}\int_{-\infty}^{\infty}\frac{1}{b}\left(\frac{1}{2}g''(\frac{z_1}{b})(\frac{\bar{s}-\bar{u}}{b})^2\right)dz = \frac{1}{|S_A|}\sum_{s\in S_A} O\left(\frac{(\bar{s}-\bar{u})^2}{b^2}\right) = O\left(\frac{\mathbb{E}_{s\in S_A}(\bar{s}-\bar{u})^2}{b^2}\right) \tag{A.13}$$

which proves the theorem.