Adaptive Multi-Hierarchical signSGD for Communication-Efficient Distributed Optimization

Haibo Yang[†] Xin Zhang[‡†] Minghong Fang[†] Jia Liu[†]
[†]Department of Computer Science, Iowa State University
[‡]Department of Statistics, Iowa State University

***Abstract*—In this work, we investigate a communication-efficient multi-hierarchical signSGD (MH-signSGD) algorithm with an adaptive learning rate. Under the symmetric assumption of the stochastic gradient distribution, we show that, without the need for learning rate tuning, our proposed MH-signSGD matches the state-of-art sublinear convergence rate $O(1/\sqrt{K})$ in nonconvex settings, where K is the number of iterations. Our adaptive learning strategy is based on stochastically approximating the learning rate found by greedily minimizing an error upper bound between two successive iterations. Moreover, by leveraging a normal approximation technique to characterize stochastic gradient sign error, we are able to sharpen the convergence analysis of MH-sighSGD with a fixed learning rate $1/\sqrt{K}$ and establish a strong result in the large-system regime, which says that the MH-signSGD algorithm asymptotically converges to a stationary point at rate $O(1/\sqrt{M})$, where M is the number of workers. In comparison, most existing work on signSGD can only prove a weaker finite neighborhood convergence in the large system regime. We validate our theoretical results experimentally both on synthetic data and real-world datasets.**

***Index Terms*—Distributed optimization, communication-efficiency, adaptive learning rate, stochastic gradient descent.**

I. Introduction

In recent years, machine learning (ML) and artificial intelligence (AI) rapidly emerge as key enabling technologies that fundamentally change our everyday life. At the heart of the training phase of many ML/AI applications lies the problem of empirical risk minimization (ERM), which can be written in the form of $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x},\zeta_i)$, where the vector $\mathbf{x}$ contains the training model parameters, ζ_i represents the i-th training sample, $f_i(\cdot)$ represents a loss function that measures the difference between the model output and the ground truth corresponding to sample ζ_i, and N is the total size of the training dataset. To date, the standard training algorithm in most ML/AI applications remains the basic stochastic gradient method (SGD), which is due to its low implementation complexity. However, as most ML/AI trainings increasingly rely on big data (which implies large N), the standard SGD method is time-consuming due to its inherent sequential nature. To address this challange, a viable solution is to exploit the massive parallelism in distributed computing to implement SGD, as evidenced in modern GPU servers or even large-scale data centers.

Although distributed SGD has been widely adopted, two fundamental challanges arise that could affect the future prospects of ML/AI. The first challenge is the high communication cost when the problem dimension d and dataset size N are large. To alleviate the high communication cost, there has been a growing interest in studying various gradient compression schemes for distributed SGD [1]. One notable example is the so-called signSGD method proposed by Bernstein *et al.* [2], where the basic idea is to take the element-wise signs of stochastic gradient vectors. In essence, signSGD can be viewed as taking one-bit quantization in each gradient coordinate. Moreover, for the distributed computing setting, Bernstein *et al.* [2] further proposed a multi-hierarchical signSGD approach (called MH-signSGD in this paper for short)[1]. It was shown in [2] that MH-signSGD converges with a sublinear rate $O(1/\sqrt{K})$, where K is the maximum iterations the algorithm runs. However, to achieve this convergence rate, the batch size should also be chosen as $O(K)$, which could be unrealistic in practice when K is large. In their follow-up work [3], they further proved the convergence of signSGD with smaller batch sizes not dependent on K.[2] However, it is unclear whether this improved batch size result can be extended to MH-signSGD or not. Moreover, the convergence measure is based on a combination of ℓ^1- and ℓ^2-norms, which makes direct convergence comparisons to other SGD-based algorithms inconvenient. These limitations necessitate a new convergence analysis for MH-signSGD.

Another limitation of SGD-based algorithms is that they are highly sensitive to the choice of learning rates, which may require a significant amount of efforts to fine-tune. One popular approach to address this challenge is based on adaptive learning rates (e.g., Adam [4], etc.), which utilizes historical stochastic gradients information to adaptively adjust learning rates. Unfortunately, it is well-known that the convergence of SGD-based methods with such adaptive learning rates may not be guaranteed. As a result, there is a compelling need to explore alternative adaptive learning rate approaches that are not based on historical stochastic gradient information. This, to our knowledge, remains an open problem. In addition to convergence analysis and adaptive learning rates, a fundamental but overlooked question is how MH-signSGD performs in the large system regime as the number of workers increases asymptotically. In other words, it is not well-understood whether the increased computing resources from a larger number of workers always lead to better convergence.

The above limitations of the existing work on MH-signSGD

[1]It was also named "signSGD by majority vote" in [2].
[2]The batch size could be as small as one.

motivate us to conduct a deeper analysis for MH-signSGD with a new adaptive learning rate design. Our main contributions are summarized as follows:

- We show that, under the symmetric assumption of the stochastic gradient distribution, the MH-signSGD method achieves a sublinear $O(1/\sqrt{K})$ convergence rate in nonconvex settings, where K is the number of iterations. Moreover, this convergence rate is achieved under arbitrary constant batch sizes independent of K. In addition, we evaluate the convergence rate in ℓ^2-norm rather than ℓ^1-norm [2] or a mixture of ℓ^1- and ℓ^2 -norms [3], which facilitates direct comparisons to other SGD-based methods.
- By leveraging a normal approximation technique to characterize the error of stochastic gradient signs, we are able to sharpen the convergence analysis and establish a strong result in the large-system regime, which states that the MH-signSGD algorithm asymptotically converges to a stationary point at rate $O(1/\sqrt{M})$, where M is the number of workers. This result shows that, under any finite maximum number of iterations, the MH-signSGD method can converge arbitrarily close to a stationary point as the number of workers increases asymptotically. In comparison, most existing work on signSGD can only prove a weaker finite neighborhood convergence result in the large-system regime. This new result advances our understanding of MH-signSGD.
- We propose a new adaptive learning rate strategy from a statistical perspective. Specifically, our adaptive learning rate strategy is based on stochastically approximating the learning rate found by greedily minimizing an error upper bound between two successive iterations. To our knowledge, this is the first adaptive learning rate strategy based on local sampling for first-order methods, which is different from conventional adaptive methods using historical information of the stochastic gradients. In addition, we prove a sufficient condition for any adaptive learning rate to converge, which could be used to evaluate the convergence of other adaptive methods and hence of independent interest.

The remainder of this paper is organized as follows: Section II focuses on the MH-signSGD algorithmic design and performance analysis. Section III presents numerical results and Section IV concludes this paper. Proof details are provided in [5] due to space limitation.

Notation: In this paper, we use boldface to denote matrices/vectors. We let $[\cdot]_i$ represent the i-th entry of a vector. We use $\|\cdot\|_2$ to denote the ℓ^2-norm. The operator $\mathtt{sign}(\cdot)$ on vectors is to be understood as taking signs element-wise. We use $[K]$ to denote the set of integers $\{1, 2, \ldots, K\}$ for any natural number K.

II. Algorithm and Convergence Analysis

In this section, we will first introduce the multi-hierarchical signSGD (MH-signSGD) algorithm in Section II-A. Then, we will present the convergence analysis and the design of adaptive learning rates (and the associated convergence analysis) for the MH-signSGD method in Section II-B. Proof sketches of the main theoretical results are give in Section II-C.

A. The MH-signSGD Algorithm

The distributed MH-signSGD algorithm is presented in Algorithm 1, where we first choose appropriate hyperparameters, e.g., learning rate, maximum number of iterations, etc. In each iteration, each worker returns the element-wise signs of the current stochastic gradient (could be based on a single sample or a minibatch). At the parameter server side, upon the receptions of all sign information from the workers, the parameter server takes another sign operation of the aggregated signs and updates the model parameters using the aggregated sign, hence the name "multi-hierarchical signSGD."

Algorithm 1 The distributed MH-signSGD method.

1: **Input:** Learning rate $\{\eta_k\}$, worker number M, initial parameter $\{\mathbf{x}_0\}$, maximum number of iterations K.
2: **for** $k = 0, 1, ..., K-1$ **do**
3: **On each worker:**
$\mathtt{sign}\{\widetilde{\mathbf{g}}_{k,j}\} \leftarrow$ Sign of gradient returned by the jth worker ($j \in [M]$) at iteration k
4: **On parameter server (PS):**
$\hat{\mathbf{g}}_k = \mathtt{sign}\{\sum_{j=1}^{M} \mathtt{sign}(\widetilde{\mathbf{g}}_{k,j})\}$
$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \hat{\mathbf{g}}_k$
5: **end for**

We note that the idea of MH-signSGD can also be implemented in a centralized fashion as shown in Algorithm 2 if we interpret the number of workers in the distributed setting as the size of minibatch in a centralized setting (note that the only difference is that the batch size parameter m in Algorithm 2 replaces the number of workers M in Algorithm 1).

Algorithm 2 The centralized MH-signSGD method.

1: **Input:** Learning rate $\{\eta_k\}$, batch size m, initial parameter $\{\mathbf{x}_0\}$, maximum number of iterations K.
2: **for** $k = 0, 1, ..., K-1$ **do**
3: $\widetilde{\mathbf{g}}_{k,j} \leftarrow$ Stochastic gradient of the jth data sample ($j \in [m]$) at iteration k
4: $\hat{\mathbf{g}}_k = \mathtt{sign}\{\sum_{j=1}^{m} \mathtt{sign}(\widetilde{\mathbf{g}}_{k,j})\}$
5: $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \hat{\mathbf{g}}_k$
6: **end for**

In what follows, we will present our main theoretical results for the MH-signSGD method.

B. Main Theoretical Results

We first state the assumptions that are needed for later results:

Assumption 1 (Lower Boundedness)**.** *For all $\mathbf{x} \in \mathbb{R}^d$, there exists f^* such that $f(\mathbf{x}) \geq f^*$.*

Assumption 2 (Unbiased Gradient Estimator)**.** *The stochastic gradient $\widetilde{\mathbf{g}}(\mathbf{x})$ satisfies $\mathbb{E}[\widetilde{\mathbf{g}}(\mathbf{x})] = \mathbf{g}(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^d$.*

Assumption 3 (Lipschitz continuous gradient)**.** *There exists a constant $L > 0$, such that $\forall \mathbf{x}, \mathbf{y}$, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall i$.*

Assumption 4 (Symmetric Gradient Noise)**.** *For every data sample (or a mini-batch data), the stochastic gradient* $\widetilde{\mathbf{g}}(\mathbf{x})$ *has a symmetric distribution around the mean.*

Assumptions 1-3 are standard assumptions used in convergence analysis for SGD algorithms (e.g., [6], [7]). Assumption 2 and 4 directly imply $\mathbb{P}[\mathtt{sign}(\widetilde{\mathbf{g}}(\mathbf{x})) = \mathtt{sign}(\mathbf{g}(\mathbf{x}))] \geq \frac{1}{2}$, $\forall \mathbf{x}$. We note that the symmetry assumption is reasonable in many practical cases where the samples are i.i.d., for which the Central Limit Theorem (CLT) suggests that the gradient noise distribution under a reasonable batch size is near Gaussian [2], [3][3]. Empirical studies [10] have also found that gradient noises are highly likely distributed as symmetric α-stable distributions even in situations where CLT may not applicable.

Under these assumptions, our first main result is on the convergence rate of the distributed MH-sign-SGD method:

Theorem 1 (Convergence of Distributed MH-signSGD)**.** *In a distributed system with* M *workers, under Assumptions 1–4, the sequence of output* $\{\mathbf{x}_k\}$ *generated by Algorithm 1 with a fixed learning rate* $\eta_k = \frac{1}{\sqrt{K}}$ *satisfies:*

$$\min_{k\in[K]} \mathbb{E}\|\mathbf{g}(\mathbf{x}_k)\|_2^2 \lessapprox \frac{1}{c\sqrt{K}}\left(f_0 - f_* + \frac{L}{2}\right),$$

where $c = b_1\sqrt{M} + b_2 > 0$*, and* $b_1 > 0, b_2 < 0$ *are constants.*

Remark 1. Three remarks for Theroem 1 are in order: 1) Theorem 1 implies that, under the symmetric assumption of the stochastic gradient distribution, the MH-signSGD method converges to a stationary point at a sublinear rate $O(1/\sqrt{K})$, which matches the state of the art. Also, the convergence rate is characterized by the conventional ℓ^2-norm, which allows direct comparisons with other SGD-based methods. 2) The $O(1/\sqrt{K})$ convergence rate is achieved by *any arbitrary constant batch size*. By contrast, the existing works on signSGD [2], [3] either require batch size K or a smaller batch size that is only limited to the centralized setting. 3) Note also that Theorem 1 implies that the convergence to a stationary point can be arbitrarily close in the large-system regime, where the number of workers increases asymptotically. This is a consequence of the use of a new proof technique based on normal approximation. This result is *stronger* than the existing works [2], [3], where only a convergence to a stationary point neighborhood can be proved.

Recall that by viewing the number of workers M as the batch size m, the MH-signSGD method can be implemented in the centralized setting straightforwardly. Hence, the result below for the centralized setting immediately follows:

Corollary 2 (Convergence of Centralized MH-signSGD)**.** *In a centralized system with batch size* m *and under Assumptions 1–4, the sequence of output* $\{\mathbf{x}_k\}$ *generated by Algorithm 1 with a fixed learning rate* $\eta_k = \frac{1}{\sqrt{K}}$ *satisfies:*

$$\min_{k\in[K]} \mathbb{E}\|\mathbf{g}(\mathbf{x}_k)\|_2^2 \lessapprox \frac{1}{c\sqrt{K}}\left(f_0 - f_* + \frac{L}{2}\right),$$

[3]The Gaussian gradient noise assumption has also been adopted in several studies for SGD, see, e.g., [8], [9].

where $c = b_1\sqrt{m} + b_2 > 0$*, and* $b_1 > 0, b_2 < 0$ *are constants.*

Remark 2. Note that in the centralized setting, if we adaptively choose the batch size $m = \Theta(K)$, then Corollary 2 implies an $O(1/K)$ convergence rate, which matches the convergence rate of the standard gradient descent method.

Notice that in Theorem 1, Corollary 2, and existing work of signSGD [2], [3], the learning rate is chosen as $1/\sqrt{K}$, which depends on the maximum number of iterations. In practice, K is usually large, which implies an extremely conservative learning rate and slow convergence. This motivates us to consider the design of adaptive learning rate to address this limitation. Toward this end, we first establish a sufficient condition that guides the development of our adaptive learning rate strategy:

Proposition 3 (Sufficient Condition of Learning Rate)**.** *In each iteration, if the learning rate is in the interval* $(0, 2\eta_k^*)$*, where*

$$\eta_k^* = \frac{\sum_{i=1}^d |[\mathbf{g}_k]_i|(1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)])}{L}, \quad (1)$$

then the sequence $\{f_k\}$ *produced by Algorithm 1 converges.*

Remark 3. The proof of Propostion 3 is based on the following observation in the proof of Theorem 1:

$$\mathbb{E}[f_{k+1} - f_k|\zeta_k] \leq \frac{L}{2}\eta_k^2 - \eta_k \sum_{i=1}^d |[\mathbf{g}_k]_i|(1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)]). \quad (2)$$

If the learning rate is chosen in such a way that the right-hand side (RHS) is negative, a decrement of f is guaranteed in each iteration. Then, greedily minimizing the upper bound on the RHS in each iteration (a quadratic function of η_k) yields the stated result in Proposition 3. This offers new opportunities in developing an adaptive learning rate strategy.

Proposition 3 provides a hint that when η_k^* is adopted, the upper bound of average decrement $\mathbb{E}[f_{k+1} - f_k|\zeta_k]$ is minimized. This gives rise to a potential learning rate choice provided that the sign error probability and the problem's smoothness parameter L are available or can be accurately estimated. Although having these conditions, adopting learning rate based on η_k^* could be more advantageous compared to the learning rate choice in SGD. In SGD, a good learning rate often relies on the knowledge of variance $\mathbb{E}[\|\widetilde{\mathbf{g}}_k\|_2^2]$, which may not be available in practice. By contrast, the η_k^* in Proposition 3 does not rely on stochastic gradient variance explicitly. The following result characterizes the convergence performance when the idealized learning rate η_k^* is adopted.

Theorem 4 (Greedy-Based Adaptive Learning Rate)**.** *Under Assumptions 1–4, the output* $\{\mathbf{x}_k\}$ *generated by Algorithm 1 with an greedy-based adaptive learning rate* $\eta_k = \eta_k^*$ *satisfies:*

$$\min_{k\in[K]} \|\mathbf{g}_k\|_2^2 \lessapprox \frac{1}{c\sqrt{K}}\sqrt{2L(f_0 - f_*)},$$

where $c = b_1\sqrt{M} + b_2 > 0$*, and* $b_1 > 0, b_2 < 0$ *are constants.*

Remark 4. With the greedy-based adaptive learning η_k^*, the convergence rate of MH-signSGD is $O(1\sqrt{K})$, i.e., the convergence rate remains the same in order sense. Adopting η_k^* as the learning rate can be viewed as an adaptive approach due to the following reasons: in the initial stage where the magnitude of stochastic gradient norm dominates the gradient variance, the sign error probability $\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]$ is small and hence a large learning rate (cf. Eq. (1)). On the other hand, when the iterates approach a stationary point, the variance of stochastic gradients is no longer negligible compared to the norm of stochastic gradients (or even dominates the norm), the sign error probability $\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]$ is large, which implies a small learning rate (cf. Eq. (1)).

Note that the rationale behind our adaptive learning rate approach is quite different from those of the conventional adaptive counterparts (e.g., Adam [4] and many others [11]), most of which rely on past stochastic gradient information.[4] Our adaptive strategy can be viewed as a statistical approach that is only based on the current stochastic gradient information (see the $\widetilde{\mathbf{g}}_k$-terms in Eq. (1)), thus eliminating the need for intricate manipulations of the relationship between current and past stochastic gradients. To our knowledge, this work is the first statistical adaptive learning rate approach.

Note that to adopt η_k^* as the learning rate, it remains to know the sign error probability $\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]$ and the Lipschitz constant L. So far, there has been a vast body of work on estimating L (e.g., [13]). To estimate the sign error probability, one can use the empirical average of the stochastic gradients in a minibatch to approximate the true gradient $\mathbf{g}_k$ and count the number of sign mismatches between $\widetilde{\mathbf{g}}_k$ and estimated true gradient.

C. Proofs of the Main Results

Due to space limitation, we provide proof sketches for the main results in this section and relegate the full proofs to [5].

Proof Sketch for Theorem 1. The first step of the proof is to bound the per-iteration decrement of the objective function $f(\cdot)$, for which we have:

$$\mathbb{E}[f_{k+1} - f_k|\zeta_k] \leq \frac{L}{2}\eta_k^2 - \eta_k \times \sum_{i=1}^{d} |[\mathbf{g}_k]_i| \left(1 - 2\mathbb{P}\left[\texttt{sign}\left(\sum_{j=1}^{m} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)\right) \neq \texttt{sign}([\mathbf{g}_k]_i)\right]\right).$$

After taking full expectation, telescoping, and rearranging, we can derive that:

$$\min_{k\in[K]} \left\{ \mathbb{E}\left[\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}\left[\texttt{sign}(\sum_{j=1}^{m} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)) \right] \neq \texttt{sign}([\mathbf{g}_k]_i)\right]) \right\} \leq \frac{f_0 - f_* + \frac{L}{2}\sum_{k=0}^{K-1}\eta_k^2}{K}. \quad (3)$$

[4]The effectiveness of using historical gradient information for adaptive learning rates and exploring modifications to such adaptive methods remains an active research field [12].

Note that the most challenging part in simplifying the left-hand-side is the product of the probability and norm terms. Toward this end, we use a normal approximation technique due to CLT to bound the probability term as follows:

$$\left(1 - 2\mathbb{P}\left[\texttt{sign}\left(\sum_{j=1}^{m} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)\right)\right] \right) \neq \texttt{sign}([\mathbf{g}_k]_i)]) \geq c[\mathbf{g}_k]_i.$$

Plugging the above results into (3) and setting $\eta_k = 1/\sqrt{K}$ yields the final results and the proof is complete. □

Proof Sketch for Proposition 3. We again start by analyzing the per-iteration decrement of the objective function and obtain Eq. (2). By choosing $\eta_k \in (0, 2\eta_k^*)$, where η_k^* is defined in (1), the RHS of Eq. (2) is negative, which means the sequence $\{f_k\}$ is monotonically decreasing. Note that f_k is bounded from below by Assumption 1. Then, the convergence with $\eta_k \in (0, 2\eta_k^*)$ follows from the Monotone Convergence Theorem. □

Proof Sketch for Theorem 4. Consider the per-iteration decrement in Eq. (2). Once the stochastic gradient $\widetilde{\mathbf{g}}_k]$ is given, the upper bound in the RHS of Eq. (2) reduces to a quadratic function of the learning rate η_k. Therefore, minimizing the upper bound amounts to solving a quadratic minimization problem as follows:

$$\min_{\eta_k} \left\{ -\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]) + \frac{L}{2}\eta_k^2 \right\}.$$

The solution to the above problem is exactly the η_k^* as shown in Eq. (1). Plugging η_k^* into above equation, we have:

$$\min_{\eta_k} \left\{ -\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]) + \frac{L}{2}\eta_k^2 \right\} = \frac{[\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)])]^2}{2L}.$$

By taking expectation, telescoping, and rearranging, we have:

$$\min_{k\in[K]} \mathbb{E}\left\{ \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]) \right\}^2 \leq \frac{2L(f_0 - f_*)}{K}. \quad (4)$$

Using the same normal approximation technique as in the proof of Theorem 1, we have:

$$\mathbb{E}[\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)])]^2 \gtrapprox \mathbb{E}\{\sum_{i=1}^{d} c|[\mathbf{g}_k]_i|^2\}^2 \geq \mathbb{E}\sum_{k=1}^{K} c^2 \|\mathbf{g}_k\|_2^4. \quad (5)$$

Combining (4) and (5), we have

$$\min_{k\in[K]} \|\mathbf{g}_k\|_2^2 \lessapprox \frac{1}{c\sqrt{K}} \sqrt{2L(f_0 - f_*)}.$$

This completes the proof. □

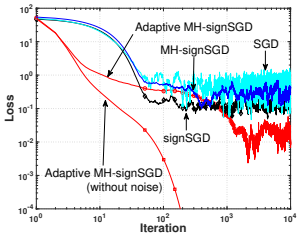

Fig. 1. Comparisons of SGD, signSGD and MH-signSGD with synthetic data.

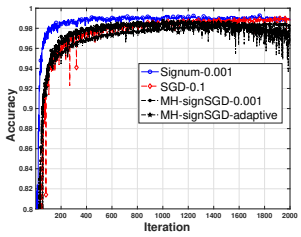

Fig. 2. Comparisons of SGD, signSGD and MH-signSGD on the MNIST dataset.

III. Numerical Results

In this section, we conduct numerical experiments to evaluate the performance of MH-signSGD with various learning rate choices and compare them with those of the existing work.

1) Synthetic Data: We first evaluate the convergence performance of MH-signSGD using synthetic data. For fair comparisons, we adopt the same problem setting as in [2], [14], where the objective function is $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^2$. Gaussian noise drawn from $\mathbb{N}(0, 100^2)$ is added to the first component of the gradient. Here, we compare MH-signSGD with SGD and signSGD with the same learning rate 0.05. We also test MH-signSGD with the proposed adaptive approach. The results are shown in Fig. 1, where each curve is averaged over 10 trials with batch size 256. From Fig. 1, we can see that MH-signSGD, signSGD, and SGD perform similarly under the same fixed learning rate. In comparison, the training loss of MH-signSGD with the proposed adaptive learning rate is approximately one order of magnitude better after 10^4 iterations. To understand our adaptive learning rate, we also plot the ideal case where the noise of gradient is zero. We can see that MH-signSGD with adaptive learning rates closely tracks the curve of the ideal case in the initial stage. This confirms our intuition that, in the initial stage where the gradients noise negligible compared to its magnitude, the sign error probability is small (cf. Remark 4).

2) Real-World Data: Next, we perform the same set of comparisons on the MNIST dataset using convolutional neural networks (CNN). The architecture of CNN has two convolution layers (size 5×5), each of which is followed by a max polling layer over a 2×2 area with stride two, and a fully-connected layer with 512 units. The ReLU activation function is used for all layers. The batch size for all algorithms is 256. For MH-signSGD, we use four workers, each of which has a mini-batch size 64. For the Lipschitz constant in adaptive MH-signSGD, we use the dimension of the parameter vector as an approximation. We compare the learning accuracy performances of MH-signSGD with SGD and Signum (the momentum version of signSGD). Comprehensive learning rate choices from 10^{-5} to 0.1 are explored and the best learning rate results are shown in Fig. 2. We can observe that all algorithms achieve similar learning accuracy. It is worth pointing out that our adaptive MH-signSGD achieves this learning accuracy without tuning the learning rate, while the other algorithms require significant efforts in identifying good learning rates.

IV. Conclusion

In this paper, we considered a multi-hierarchical signSGD (MH-signSGD) algorithm, with the goal to achieve both easy learning rate selection and communication efficiency for distributed optimization. With a fixed learning rate and under the symmetric assumption of the stochastic gradient distribution, we proved a stronger $O(1/\sqrt{K})$ result in the sense that the convergence of MH-signSGD can be arbitrarily close to a stationary point when the number of workers increases asymptotically in the large-system regime. We further developed a new adaptive learning rate strategy based on stochastically approximating the learning rate found by greedily minimizing an error upper bound between two successive iterations. Our approach does not require any intricate manipulations of the relationships between current and past stochastic gradients. We conducted extensive numerical studies based on synthetic data and real-world datasets, and the numerical results confirmed our theoretical findings.

References

[1] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," vol. 80. PMLR, 2018, pp. 560–569.

[3] J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar, "signSGD with majority vote is communication efficient and fault tolerant," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=BJxhijAcY7

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[5] H. Yang, X. Zhang, M. Fang, and J. Liu, "Adaptive multi-hierarchical signSGD for communication-efficient distributed optimization," Iowa State University, Tech. Rep., February 2020. [Online]. Available: http://web.cs.iastate.edu/~jialiu/publications/MH_signSGD_TR.pdf

[6] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[7] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[8] W. Hu, C. J. Li, L. Li, and J.-G. Liu, "On the diffusion approximation of nonconvex stochastic gradient descent," *Annals of Mathematical Sciences and Applications*, vol. 4, no. 1, pp. 3–32, 2019.

[9] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, "Three factors influencing minima in sgd," *arXiv preprint arXiv:1711.04623*, 2017.

[10] U. Simsekli, L. Sagun, and M. Gurbuzbalaban, "A tail-index analysis of stochastic gradient noise in deep neural networks," *arXiv preprint arXiv:1901.06053*, 2019.

[11] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[12] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ

[13] G. Wood and B. Zhang, "Estimation of the lipschitz constant of a function," *Journal of Global Optimization*, vol. 8, no. 1, pp. 91–103, 1996.

[14] S. Liu, P.-Y. Chen, X. Chen, and M. Hong, "signSGD via zeroth-order oracle," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=BJe-DsC5Fm

[15] P. Hansen, "Approximating the binomial distribution by the normal distribution–error and accuracy," 2011.

APPENDIX A
PROOF FOR THEOREM 1

Due to Lipschitz continuity of the gradients (Assumption 3) and iteration step in Algorithm 1:

$$
\begin{aligned}
&f_{k+1} - f_k \leq \mathbf{g}_k^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \sum_{i=1}^{d} \frac{L_i}{2}[(\mathbf{x}_{k+1} - \mathbf{x}_k)]_i^2 \\
&= -\eta_k \mathbf{g}_k^T \texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}])) + \eta_k^2 \sum_{i=1}^{d} \frac{L_i}{2} \\
&= -\eta_k \|\mathbf{g}_k\|_1 + \eta_k^2 \sum_{i=1}^{d} \frac{L_i}{2} + \\
&2\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| \mathbb{I}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq \texttt{sign}([\mathbf{g}_k]_i)]
\end{aligned}
$$

Let $L = \sum_{i=1}^{d} L_i$

$$
\begin{aligned}
&\mathbb{E}[f_{k+1} - f_k | \zeta_k] \leq -\eta_k \|\mathbf{g}_k\|_1 + \\
&2\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| \mathbb{P}[sign(\sum_{j=1}^{M} sign([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq sign([\mathbf{g}_k]_i)] + \frac{L}{2}\eta_k^2 \\
&= \frac{L}{2}\eta_k^2 - \\
&\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq \texttt{sign}([\mathbf{g}_k]_i)])
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
&\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq sign([\mathbf{g}_k]_i)]) \\
&\leq \mathbb{E}[f_k - f_{k+1} | \zeta_k] + \frac{L}{2}\eta_k^2
\end{aligned}
$$

Taking expection and telescoping:

$$
\begin{aligned}
&\mathbb{E} \sum_{k=0}^{K-1} \eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - \\
&2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} sign([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq \texttt{sign}([\mathbf{g}_k]_i)]) \\
&\leq f_0 - f_* + \frac{L}{2} \sum_{k=0}^{K-1} \eta_k^2
\end{aligned}
$$

That is:

$$
\begin{aligned}
&\min_{k \in [K]} \{\mathbb{E}[\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - \\
&2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i))] \neq \texttt{sign}([\mathbf{g}_k]_i)])\} \\
&\leq \frac{f_0 - f_* + \frac{L}{2} \sum_{k=0}^{K-1} \eta_k^2}{K} \qquad (*)
\end{aligned}
$$

Qualitatively, as $K \to \infty$ and $\sum_{k=1}^{K} \eta_k^2 < \infty$, which means $\frac{f_0 - f_* + \frac{L}{2} \sum_{k=1}^{K} \eta_k^2}{K} \to 0$. Then,

$$
\begin{aligned}
&\min\{\mathbb{E}[\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - \\
&2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i))] \neq \texttt{sign}([\mathbf{g}_k]_i)])\} \to 0
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
&\min\{\mathbb{E}[\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - \\
&2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i))] \neq \texttt{sign}([\mathbf{g}_k]_i)])\} \to 0
\end{aligned}
$$

$\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} \texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i)) \neq \texttt{sign}([\mathbf{g}_k]_i)]) \to \frac{1}{2}$ if and only if $[\mathbf{g}_k]_i \to 0$ due to Assumption 4. Hence, $\min\{[\mathbf{g}_k]_i\} \to 0$. Now, the most challenging part is to handle the entanglement of probability term and norm term. Here we use normal approximation of CLT.

Let z_i be the number of $[\widetilde{\mathbf{g}}_{k,j}]_i$ such that $\texttt{sign}([\widetilde{\mathbf{g}}_{k,j}]_i) \neq \texttt{sign}[\mathbf{g}_k]_i, j \in [M]$, then $z_i \sim$ Binomial(M, p), here $p = p(k,i) \in (0, \frac{1}{2})$ due to assumption 4 and we assume $[\mathbf{g}_k]_i > 0$ without loss of generality.

$$
\begin{aligned}
&(1 - 2\mathbb{P}[\texttt{sign}(\sum_{j=1}^{M} sign([\widetilde{\mathbf{g}}_{k,j}]_i))]_i) \neq \texttt{sign}([\mathbf{g}_k]_i)]) \\
&= 1 - 2\mathbb{P}[z_i > \frac{M}{2}] \\
&= 2\mathbb{P}[z_i \leq \frac{M}{2}] - 1 \\
&= 2\Phi(\frac{\frac{M}{2} - Mp}{\sqrt{Mp(1-p)}}) + 2[F(\frac{M}{2}) - \Phi(\frac{\frac{M}{2} - Mp}{\sqrt{Mp(1-p)}})] - 1 \\
&\overset{(a)}{\geq} 2\Phi(\frac{\frac{M}{2} - Mp}{\sqrt{Mp(1-p)}}) - \epsilon_{abs}(p) - 1 \\
&= \underbrace{\mathbf{erf}(x)}_{(1)} - \underbrace{\epsilon_{abs}(p)}_{(2)}
\end{aligned}
$$

Inequality (a) is due to normal approximation for Binomial distribution. Here, $\mathbf{erf}$ is the error function and $x = \frac{\frac{M}{2} - Mp}{\sqrt{2Mp(1-p)}}$. Φ is the CDF of normal distribution, and F is the CDF of Binomial distribution. ϵ_{abs} is the maximum absolute error term, i.e., $\epsilon_{abs}(p) = max\{|F(k,p) - \Phi(\frac{k - Mp}{\sqrt{Mp(1-p)}})|\}$. We view ϵ_{abs} as a function of p when k is fixed and large enough.

When $g_k \to 0$ then $p \to \frac{1}{2}$, $x \to 0$. We have $\frac{1}{2} - p \approx a_1 [\mathbf{g}_k]_i$ and $x = \frac{\frac{M}{2} - Mp}{\sqrt{2Mp(1-p)}} = \frac{M(\frac{1}{2} - p)}{\sqrt{2Mp(1-p)}} \approx \frac{M}{\sqrt{2Mp(1-p)}} a_1 [\mathbf{g}_k]_i$, where $a_1 > 0$ is the first order taylor expansion coefficient of $p([\mathbf{g}_k]_i)$.

For term (1): $\mathbf{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{i=0}^{\infty} \frac{(-1)^i x^{2i+1}}{i!(2i+1)}$. Then $\mathbf{erf}(x) \geq \frac{2}{\sqrt{\pi}}(x - \frac{1}{6}x^3)$ since function $y = \frac{2}{\sqrt{\pi}}(\frac{1}{\sqrt{2}}x - \frac{1}{6\sqrt{2}}x^3)$ is a concave function when $x \in (0,1)$, so $\mathbf{erf}(x) \geq a_2 x \approx a_1 a_2 \sqrt{m} \frac{1}{\sqrt{2p(1-p)}} [\mathbf{g}_k]_i$ where $a_2 = \frac{5}{3\sqrt{2\pi}}$.

For term (2): $\epsilon_{abs}(p) = max\{|F(k,p) - \Phi(\frac{k-Mp}{\sqrt{Mp(1-p)}})|\} = max\{|(\frac{1}{2} - F(k)) - (\frac{1}{2} - \Phi(\frac{k-Mp}{\sqrt{Mp(1-p)}}))|\}$, both CDF of binomial and normal distribution are analytic functions. Using Taylor's expansion, we can bound $\epsilon_{abs}(p) \leq c_1(\frac{1}{2} - p)$, where c_1 is some positive constant. So $\epsilon_{abs}(p) \leq c_1(\frac{1}{2} - p) \lessapprox c_1 a_1 [\mathbf{g}_k]_i$. Empirical investgation in statistics also support this result [15].

Then we have:

$$\begin{aligned}\mathbf{erf}(x) - \epsilon_{abs}(p) &\gtrapprox a_1 a_2 \sqrt{M} \frac{1}{\sqrt{2p(1-p)}} [\mathbf{g}_k]_i - c_1 a_1 [\mathbf{g}_k]_i \\ &= (a_1 a_2 \sqrt{M} \frac{1}{\sqrt{2p(1-p)}} - c_1 a_1)[\mathbf{g}_k]_i \\ &= c[\mathbf{g}_k]_i\end{aligned}$$

where $c = a_1 a_2 \sqrt{M} \frac{1}{\sqrt{2p(1-p)}} - c_1 a_1 = b_1 \sqrt{M} + b_2 > 0, b_1 = a_1 a_2 \frac{1}{\sqrt{2p(1-p)}}, b_2 = -c_1 a_1$

Plugging into $(*)$ and letting $\eta_k = \frac{1}{\sqrt{K}}$, we have

$$\min_{k \in [K]} \{\mathbb{E}[\eta_k \sum_{i=1}^{d} [\mathbf{g}_k]_i^2]\} \lessapprox \frac{1}{cK}(f_0 - f_* + \frac{L}{2} \sum_{k=1}^{K} \eta_k^2)$$

i.e.,

$$\min_{k \in [K]} \{\mathbb{E}[\|g_k\|_2^2]\} \lessapprox \frac{1}{c\sqrt{K}}(f_0 - f_* + \frac{L}{2} \sum_{k=1}^{K} \eta_k^2)$$

APPENDIX B
PROOF FOR PROPOSITION 4

By analyzing per-iteration decrement, we have

$$\begin{aligned}&\mathbb{E}[f_{k+1} - f_k | \zeta_k] \\ &\leq -\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)]) + \frac{L}{2} \eta_k^2.\end{aligned}$$

Once the data sample, learning rate and $\mathbf{x}_k$ are given, the right-hand side can be viewed as a quadratic function of η_k:

$$\min_{\eta_k} \{-\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)]) + \frac{L}{2} \eta_k^2\}$$

Solving this problem yields

$$\eta_k^* = \frac{\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)])}{L}.$$

Thus, we have

$$\begin{aligned}&\min_{\eta_k} \{-\eta_k \sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)]) + \frac{L}{2} \eta_k^2\} \\ &\qquad = \frac{[\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)])]^2}{2L}.\end{aligned}$$

Taking the above learning rate in each iteration and telescoping:

$$\begin{aligned}&\mathbb{E} \sum_{k=0}^{K-1} \frac{\{\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)])\}^2}{2L} \\ &\leq (f_0 - f_*) \\ &\min_{k \in [K]} \mathbb{E}\{\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[\mathtt{sign}([\widetilde{\mathbf{g}}_k]_i) \neq \mathtt{sign}([\mathbf{g}_k]_i)])\}^2 \\ &\leq \frac{2L(f_0 - f_*)}{K} \qquad (3)\end{aligned}$$

Using the same strategy of normal approximation stated before, we can bound the term:

$$\begin{aligned}&\mathbb{E}[\sum_{i=1}^{d} |[\mathbf{g}_k]_i| (1 - 2\mathbb{P}[sign([\widetilde{\mathbf{g}}_k]_i) \neq sign([\mathbf{g}_k]_i)])]^2 \\ &\gtrapprox \mathbb{E}\{\sum_{i=1}^{d} c |[\mathbf{g}_k]_i|^2\}^2 \\ &\geq \mathbb{E} \sum_{k=1}^{K} c^2 \|\mathbf{g}_k\|_2^4. \qquad (4)\end{aligned}$$

Combining (3) and (4), we get:

$$\min_{k \in [K]} \|\mathbf{g}_k\|_2^4 \lessapprox \frac{1}{c^2 K} [2L(f_0 - f_*)],$$

i.e.,

$$\min_{k \in [K]} \|\mathbf{g}_k\|_2^2 \lessapprox \frac{1}{c\sqrt{K}} \sqrt{2L(f_0 - f_*)}.$$