Stochastic Sign Descent Methods: New Algorithms and Better Theory

Mher Safaryan[1] Peter Richtárik[1,2]

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2]Moscow Institute of Physics and Technology (MIPT), Moscow, Russia

June 24, 2021

Abstract

Various gradient compression schemes have been proposed to mitigate the communication cost in distributed training of large scale machine learning models. Sign-based methods, such as signSGD [Bernstein et al., 2018], have recently been gaining popularity because of their simple compression rule and connection to adaptive gradient methods, like ADAM. In this paper, we analyze sign-based methods for non-convex optimization in three key settings: (i) standard single node, (ii) parallel with shared data and (iii) distributed with partitioned data. For single machine case, we generalize the previous analysis of signSGD relying on intuitive bounds on success probabilities and allowing even biased estimators. Furthermore, we extend the analysis to parallel setting within a parameter server framework, where exponentially fast noise reduction is guaranteed with respect to number of nodes, maintaining 1-bit compression in both directions and using small mini-batch sizes. Next, we identify a fundamental issue with signSGD to converge in distributed environment. To resolve this issue, we propose a new sign-based method, *Stochastic Sign Descent with Momentum (SSDM)*, which converges under standard bounded variance assumption with the optimal asymptotic rate. We validate several aspects of our theoretical findings with numerical experiments.

Contents

1 Introduction

One of the key factors behind the success of modern machine learning models is the availability of large amounts of training data [Bottou and Le Cun, 2003, Krizhevsky et al., 2012, Schmidhuber, 2015]. However, the state-of-the-art deep learning models deployed in industry typically rely on datasets too large to fit the memory of a single computer, and hence the training data is typically split and stored across a number of compute nodes capable of working in parallel. Training such models then amounts to solving optimization problems of the form

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

where $f_n : \mathbb{R}^d \to \mathbb{R}$ represents the *non-convex* loss of a deep learning model parameterized by $x \in \mathbb{R}^d$ associated with data stored on node n. Arguably, stochastic gradient descent (SGD) [Robbins and Monro, 1951, Vaswani et al., 2019, Qian et al., 2019] in of its many variants [Kingma and Ba, 2015, Duchi et al., 2011, Schmidt et al., 2017, Zeiler, 2012, Ghadimi and Lan, 2013] is the most popular algorithm for solving (1). In its basic implementation, all workers $n \in \{1, 2, \dots, M\}$ in parallel compute a random approximation $\hat{g}^n(x_k)$ of $\nabla f_n(x_k)$, known as the *stochastic gradient*. These approximations are then sent to a master node which performs the aggregation $\hat{g}(x_k) := \frac{1}{M}\sum_{n=1}^{M}\hat{g}^n(x_k)$. The aggregated vector is subsequently broadcast back to the nodes, each of which performs an update of the form

$$x_{k+1} = x_k - \gamma_k \hat{g}(x_k),$$

updating their local copies of the parameters of the model.

Table 1: Summary of main theoretical results obtained in this work.

	Convergence rate	Gradient norm used in theory	Weak noise assumptions	Weak dependence on smoothness	Can handle biased estimator?	Can work with small minibatch?	Can handle partitioned train data?
SGD [Ghadimi and Lan, 2013]	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	l^2 norm squared	$\mathrm{Var}[\hat{g}] \le \sigma^2$	✗ $\max\limits_{i=1}^{d} L_i$	NO	YES	YES
signSGD [Bernstein et al., 2019]	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	a mix of l^1 and l^2 squared	✗ unimodal, symmetric & $\mathrm{Var}[\hat{g}_i] \le \sigma_i^2$	✓ $\frac{1}{d}\sum\limits_{i=1}^{d} L_i$	NO	YES	NO
signSGD with M Maj.Vote [Bernstein et al., 2019]	$\mathcal{O}\left(\frac{1}{K^{1/4}}\right)$ (speedup$\sim \frac{1}{\sqrt{M}}$)	l^1 norm	✗ unimodal, symmetric & $\mathrm{Var}[\hat{g}_i] \le \sigma_i^2$	✓ $\frac{1}{d}\sum\limits_{i=1}^{d} L_i$	NO	NO	NO
Signum [Bernstein et al., 2018]	$\mathcal{O}\left(\frac{\log K}{K^{1/4}}\right)$	l^1 norm	✗ unimodal, symmetric & $\mathrm{Var}[\hat{g}_i] \le \sigma_i^2$	✓ $\frac{1}{d}\sum\limits_{i=1}^{d} L_i$	NO	NO	NO
Noisy signSGD [Chen et al., 2020]	$\mathcal{O}\left(\frac{1}{K^{1/4}}\right)$	a mix of l^1 and l^2 squared	✗ absence of noise $\hat{g} = \nabla f$	✗ $\max\limits_{n=1}^{M} L^n$	NO	NO	YES
signSGD **This work** (Thm. 5, 6)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	ρ-norm	✓ $\rho_i > \frac{1}{2}$	✓ $\frac{1}{d}\sum\limits_{i=1}^{d} L_i$	YES	YES	NO
signSGD with M Maj.Vote **This work** (Thm. 7)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ (speedup$\sim e^{-M}$)	ρ_M-norm	✓ $\rho_i > \frac{1}{2}$	✓ $\frac{1}{d}\sum\limits_{i=1}^{d} L_i$	YES	YES	NO
SSDM (Alg. 3) **This work** (Thm. 8)	$\mathcal{O}\left(\frac{1}{K^{1/4}}\right)$	l^2 norm	$\mathrm{Var}[\hat{g}] \le \sigma^2$	✓ $\frac{1}{M}\sum\limits_{n=1}^{M} L^n$	NO	YES	YES

1.1 Gradient compression

Typically, communication of the local gradient estimators $\hat{g}^n(x_k)$ to the master forms the bottleneck of such a system [Seide et al., 2014, Zhang et al., 2017, Lin et al., 2018]. In an attempt to alleviate this communication bottleneck, a number of compression schemes for gradient updates have been proposed and analyzed [Alistarh et al., 2017, Wang et al., 2018, Wen et al., 2017, Khirirat et al., 2018, Mishchenko et al., 2019]. A *compression scheme* is a (possibly randomized) mapping $Q : \mathbb{R}^d \to \mathbb{R}^d$, applied by the nodes to $\hat{g}^n(x_k)$ (and possibly also by the master to aggregated update in situations when broadcasting is expensive as well) in order to reduce the number of bits of the communicated message.

Sign-based compression. Although most of the existing theory is limited to *unbiased* compression schemes, i.e., $\mathbb{E}Q(x) = x$, *biased* schemes such as those based on communicating signs of the update entries only often perform much better [Seide et al., 2014, Strom, 2015, Wen et al., 2017, Carlson et al., 2015, Balles and Hennig, 2018, Bernstein et al., 2018, 2019, Zaheer et al., 2018, Liu et al., 2019]. The simplest among these sign-based methods is signSGD (see Algorithm 1), whose update direction is assembled from the component-wise signs of the stochastic gradient.

Adaptive methods. While ADAM is one of the most popular *adaptive* optimization methods used in deep learning [Kingma and Ba, 2015], there are issues with its convergence [Reddi et al., 2019] and generalization [Wilson et al., 2017] properties. It was noted by Balles and Hennig [2018] that the behaviour of ADAM is similar to a momentum version of signSGD. Connection between sign-based and adaptive methods has long history, originating at least in Rprop [Riedmiller and Braun, 1993] and RMSprop [Tieleman and Hinton, 2012]. Therefore, investigating the behavior of signSGD can improve our understanding on the convergence of adaptive methods such as ADAM.

2 Contributions

We now present the main contributions of this work. Our key results are summarized in Table 1.

2.1 Single machine setup

• **2 methods for 1-node setup.** In the $M = 1$ case, we study two general classes of sign based methods for minimizing a smooth non-convex function f. The first method has the standard form[1]

$$x_{k+1} = x_k - \gamma_k \operatorname{sign} \hat{g}(x_k), \tag{2}$$

while the second has a new form not considered in the literature before:

$$x_{k+1} = \arg\min\{f(x_k), f(x_k - \gamma_k \operatorname{sign} \hat{g}(x_k))\}. \tag{3}$$

• **Key novelty.** The key novelty of our methods is in a *substantial relaxation* of the requirements that need to be imposed on the gradient estimator $\hat{g}(x_k)$ of the true gradient $\nabla f(x^k)$. In sharp contrast with existing approaches, we allow $\hat{g}(x_k)$ to be *biased*. Remarkably, we only need one additional and rather weak assumption on $\hat{g}(x_k)$ for the methods to provably converge: we require the signs of the entries of $\hat{g}(x_k)$ to be equal to the signs of the entries of $g(x^k) := \nabla f(x^k)$ with a probability strictly larger than $1/2$ (see Assumption 1). Formally, we assume the following bounds on success probabilities:

$$\operatorname{Prob}(\operatorname{sign} \hat{g}_i(x_k) = \operatorname{sign} g_i(x_k)) > \frac{1}{2} \tag{SPB}$$

for all $i \in \{1, 2, \ldots, d\}$ with $g_i(x_k) \neq 0$.

We provide three necessary conditions for our assumption to hold (see Lemma 1, 2 and 3) and show through a counterexample that a slight violation of this assumption breaks the convergence.

• **Convergence theory.** While our complexity bounds have the same $\mathcal{O}(1/\sqrt{K})$ dependence on the number of iterations, they have a *better dependence on the smoothness parameters* associated with f. Theorem 5 is the first result on signSGD for non-convex functions which does not rely on mini-batching, and which allows for step sizes independent of the total number of iterations K. Finally, Theorem 1 in [Bernstein et al., 2019] can be recovered from our general Theorem 5. Our bounds are cast in terms of a *novel norm-like function, which we call the ρ-norm*, which is a weighted l^1 norm with positive variable weights.

2.2 Parallel setting with shared data

• **Noise reduction at exponential speed.** Under the same SPB assumption, we extend our results to the *parallel setting* with arbitrary M nodes, where we also consider sign-based compression of the aggregated gradients. Considering the noise-free training as a baseline, we guarantee exponentially fast noise reduction with respect to M (see Theorem 7).

2.3 Distributed training with partition data

• **New sign-based method for distributed training.** We describe a fundamental obstacle in distributed environment, which prevents signSGD to converge. To resolve the issue, we propose a new sign-based method–*Stochastic Sign Descent with Momentum (SSDM)*; see Algorithm 3.

[1] $\operatorname{sign} g$ is applied element-wise to the entries $g_1, g_2, \ldots, g_d$ of $g \in \mathbb{R}^d$. For $t \in \mathbb{R}$ we define $\operatorname{sign} t = 1$ if $t > 0$, $\operatorname{sign} t = 0$ if $t = 0$, and $\operatorname{sign} t = -1$ if $t < 0$.

• **Key novelty.** The key novelty in our SSDM method is the notion of *stochastic sign* operator $\widetilde{\operatorname{sign}}: \mathbb{R}^d \to \mathbb{R}^d$ defined as follows:

$$\left(\widetilde{\operatorname{sign}}\, g\right)_i = \begin{cases} +1, & \text{with probability } \frac{1}{2} + \frac{1}{2}\frac{g_i}{\|g\|} \\ -1, & \text{with probability } \frac{1}{2} - \frac{1}{2}\frac{g_i}{\|g\|} \end{cases}$$

for all $i \in \{1, 2, \dots, d\}$ and $\widetilde{\operatorname{sign}}\, \mathbf{0} = \mathbf{0}$ with probability 1.

Unlike the deterministic sign operator, stochastic $\widetilde{\operatorname{sign}}$ naturally satisfies the SPB assumption and it gives an unbiased estimator with a proper scaling factor.

• **Convergence theory.** Under the standard bounded variance condition, our SSDM method guarantees the optimal asymptotic rate $\mathcal{O}(\varepsilon^{-4})$ without *error feedback* trick and communicating sign-bits only (see Theorem 8).

3 Success Probabilities and Gradient Noise

In this section we describe our key (and weak) assumption on the gradient estimator $\hat{g}(x)$, and show through a counterexample that without this assumption, signSGD can fail to converge.

3.1 Success probability bounds

Assumption 1 (SPB: Success Probability Bounds)**.** For any $x \in \mathbb{R}^d$, we have access to an independent (and *not necessarily unbiased*) estimator $\hat{g}(x)$ of the true gradient $g(x) := \nabla f(x)$ that if $g_i(x) \neq 0$, then

$$\rho_i(x) := \operatorname{Prob}\left(\operatorname{sign} \hat{g}_i(x) = \operatorname{sign} g_i(x)\right) > \frac{1}{2} \tag{4}$$

for all $x \in \mathbb{R}^d$ and all $i \in \{1, 2, \dots, d\}$.

We will refer to the probabilities ρ_i as *success probabilities.* As we will see, they play a central role in the convergence of sign based methods. Moreover, we argue that it is reasonable to require from the sign of stochastic gradient to show true gradient direction more likely than the opposite one. Extreme cases of this assumption are the absence of gradient noise, in which case $\rho_i = 1$, and an overly noisy stochastic gradient, in which case $\rho_i \approx \frac{1}{2}$.

Remark 1. *Assumption 1 can be relaxed by replacing bounds (4) with*

$$\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i(x) \cdot g_i(x)\right)\right] > 0, \quad \textit{if} \quad g_i(x) \neq 0.$$

However, if $\operatorname{sign} \hat{g}_i(x) \neq 0$ *almost surely (e.g.* $\hat{g}_i(x)$ *has continuous distribution), then these bounds are identical.*

Extension to stochastic sign oracle. Notice that we do *not* require $\hat{g}$ to be unbiased and we do *not* assume uniform boundedness of the variance, or of the second moment. This observation allows to extend existing theory to more general sign-based methods with a stochastic sign oracle. By a stochastic sign oracle we mean an oracle that takes $x_k \in \mathbb{R}^d$ as an input, and outputs a random vector $\hat{s}_k \in \mathbb{R}^d$ with entries in ±1. However, for the sake of simplicity, in the rest of the paper we will work with the signSGD formulation, i.e., we let $\hat{s}_k = \operatorname{sign} \hat{g}(x_k)$.

3.2 A counterexample to signSGD

Here we analyze a counterexample to signSGD discussed in [Karimireddy et al., 2019]. Consider the following least-squares problem with unique minimizer $x^* = (0,0)$:

$$\min_{x\in\mathbb{R}^2} f(x) = \frac{1}{2}\left[\langle a_1, x\rangle^2 + \langle a_2, x\rangle^2\right], \qquad a_1 = \left[\begin{smallmatrix} 1+\varepsilon \\ -1+\varepsilon \end{smallmatrix}\right],\ a_2 = \left[\begin{smallmatrix} -1+\varepsilon \\ 1+\varepsilon \end{smallmatrix}\right], \tag{5}$$

where $\varepsilon \in (0,1)$ and stochastic gradient $\hat{g}(x) = \nabla\langle a_i, x\rangle^2 = 2\langle a_i, x\rangle a_i$ with probabilities $1/2$ for $i = 1,2$. Let us take any point from the line $Z = \{(z_1, z_2)\colon z_1 + z_2 = 2\}$ as initial point x_0 for the algorithm and notice that $\operatorname{sign}\hat{g}(x) = \pm(1,-1)$ for any $x \in Z$. Hence, signSGD with any step-size sequence remains stuck along the line Z, whereas the problem has a unique minimizer at the origin.

In this example, Assumption 1 is violated. Indeed, notice that $\operatorname{sign}\hat{g}(x) = (-1)^i \operatorname{sign}\langle a_i, x\rangle \left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ with probabilities $1/2$ for $i = 1,2$. By $S \coloneqq \{x \in \mathbb{R}^2\colon \langle a_1, x\rangle \cdot \langle a_2, x\rangle > 0\} \neq \emptyset$ denote the open cone of points having either an acute or an obtuse angle with both a_i's. Then for any $x \in S$, the sign of the stochastic gradient is $\pm(1,-1)$ with probabilities $1/2$. Hence for any $x \in S$, we have low success probabilities:

$$\rho_i(x) = \operatorname{Prob}\left(\operatorname{sign}\hat{g}_i(x) = \operatorname{sign} g_i(x)\right) \le \frac{1}{2},\ i = 1,2.$$

So, in this case we have an entire conic region with low success probabilities, which clearly violates (4). Furthermore, if we take a point from the complement open cone $\bar{S}^c$, then the sign of stochastic gradient equals to the sign of gradient, which is perpendicular to the axis of S (thus in the next step of the iteration we get closer to S). For example, if $\langle a_1, x\rangle < 0$ and $\langle a_2, x\rangle > 0$, then $\operatorname{sign}\hat{g}(x) = (1,-1)$ with probability 1, in which case $x - \gamma \operatorname{sign}\hat{g}(x)$ gets closer to low success probability region S.

In summary, in this counterexample there is a conic region where the sign of the stochastic gradient is useless (or behaves adversarially), and for any point outside that region, moving direction (which is the opposite of the sign of gradient) leads toward that conic region.

3.3 Sufficient conditions for SPB

To motivate our SPB assumption, we compare it with 4 different conditions commonly used in the literature and show that it holds under general assumptions on gradient noise. Below, we assume that for any point $x \in \mathbb{R}^d$, we have access to an independent and unbiased estimator $\hat{g}(x)$ of the true gradient $g(x) = \nabla f(x)$.

Lemma 1 (see B.1)**.** *If for each coordinate $\hat{g}_i$ has a unimodal and symmetric distribution with variance $\sigma_i^2 = \sigma_i^2(x)$, $1 \le i \le d$ and $g_i \neq 0$, then*

$$\rho_i \ge \frac{1}{2} + \frac{1}{2}\frac{|g_i|}{|g_i| + \sqrt{3}\sigma_i} > \frac{1}{2}.$$

This is the setup used in Theorem 1 of Bernstein et al. [2019]. We recover their result as a special case using Lemma 1 (see Appendix C). Next, we replace the distribution condition by coordinate-wise strong growth condition (SGC) [Schmidt and Le Roux, 2013, Vaswani et al., 2019] and fixed mini-batch size.

Lemma 2 (see B.2)**.** *Let coordinate-wise variances $\sigma_i^2(x) \le c_i\, g_i^2(x)$ are bounded for some constants $c_i \ge 0$. Choose mini-batch size $\tau > 2\max_i c_i$ for stochastic gradient estimator. If further $g_i \neq 0$, then*

$$\rho_i \ge 1 - \frac{c_i}{\tau} > \frac{1}{2}.$$

Now we remove SGC and give an adaptive condition on mini-batch size of stochastic gradient for the SPB assumption to hold.

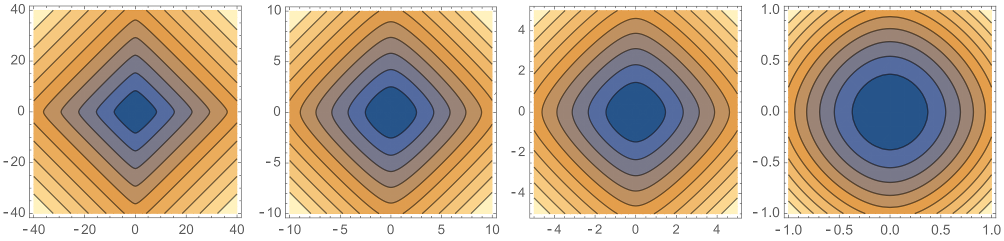

Figure 1: Contour plots of the $l^{1,2}$ norm (6) at 4 different scales with fixed noise $\sigma = 1$.

Lemma 3 (see B.3)**.** *Let $\sigma_i^2 = \sigma_i^2(x)$ be the variance and $\nu_i^3 = \nu_i^3(x)$ be the 3th central moment of $\hat{g}_i(x)$, $1 \le i \le d$. Then SPB assumption holds if mini-batch size*

$$\tau > 2 \min \left(\frac{\sigma_i^2}{g_i^2}, \frac{\nu_i^3}{|g_i|\sigma_i^2} \right).$$

Finally, we compare SPB with the standard bounded variance assumption in the sense of differential entropy.

Lemma 4 (see B.4)**.** *Differential entropy of a probability distribution under the bounded variance assumption is bounded, while under the SPB assumption it could be arbitrarily large.*

Differential entropy argument is an attempt to bridge our new SPB assumption to one of the most basic assumptions in the literature, bounded variance assumption. Clearly, they are not comparable in the usual sense, and neither one is implied by the other. Still, we propose another viewpoint to the situation and compare such conditions through the lens of information theory. Practical meaning of such observation is that SPB handles a much broader (though not necessarily more important) class of gradient noise than bounded variance condition. In other words, this gives an intuitive measure on how much restriction we put on the noise.

Note that SPB assumption describes the convergence of sign descent methods, which is known to be problem dependent (e.g. see [Balles and Hennig, 2018], section 6.2 Results). One should view the SPB condition as a criteria to problems where sign based methods are useful.

4 A New "Norm" for Measuring the Gradients

In this section we introduce the concept of a norm-like function, which we call ρ*-norm*, induced from success probabilities. Used to measure gradients in our convergence rates, ρ-norm is a technical tool enabling the analysis.

Definition 1 (ρ-norm)**.** Let $\rho \coloneqq \{\rho_i(x)\}_{i=1}^d$ be the collection of probability functions from the SPB assumption. We define the ρ-norm of gradient $g(x)$ via

$$\|g(x)\|_\rho \coloneqq \sum_{i=1}^d (2\rho_i(x) - 1)|g_i(x)|.$$

Note that mixed ρ-norm is not a norm as it may not satisfy the triangle inequality. However, under SPB assumption, it is positive definite as it is a weighted l^1 norm with positive (and variable) weights

$2\rho_i(x) - 1 > 0$. That is, $\|g\|_\rho \ge 0$, and $\|g\|_\rho = 0$ if and only if $g = 0$. Although, in general, ρ-norm is not a norm in classical sense, it can be reduced to one in special cases. For example, under the assumptions of Lemma 1, ρ-norm can be lower bounded by a mixture of the l^1 and squared l^2 norms:

$$\|g\|_\rho = \sum_{i=1}^{d}(2\rho_i - 1)|g_i| \ge \sum_{i=1}^{d} \frac{g_i^2}{|g_i| + \sqrt{3}\sigma_i} := \|g\|_{l^{1,2}}. \tag{6}$$

Note that $l^{1,2}$-norm is again not a norm. However, it is positive definite, continuous and order preserving, i.e., for any g^k, g, $\tilde{g} \in \mathbb{R}^d$ we have:

1. $\|g\|_{l^{1,2}} \ge 0$ and $\|g\|_{l^{1,2}} = 0$ if and only if $g = 0$,
2. $g^k \to g$ (in l^2 sense) implies $\|g^k\|_{l^{1,2}} \to \|g\|_{l^{1,2}}$,
3. $0 \le g_i \le \tilde{g}_i$ for any $1 \le i \le d$ implies $\|g\|_{l^{1,2}} \le \|\tilde{g}\|_{l^{1,2}}$.

From these three properties it follows that $\|g^k\|_{l^{1,2}} \to 0$ implies $g^k \to 0$. These properties are important as we will measure convergence rate in terms of the $l^{1,2}$ norm in the case of unimodal and symmetric noise assumption. To understand the nature of the $l^{1,2}$ norm, consider the following two cases when $\sigma_i(x) \le c|g_i(x)| + \tilde{c}$ for some constants c, $\tilde{c} \ge 0$. If the iterations are in ε-neighbourhood of a minimizer x^* with respect to the l^∞ norm (i.e., $\max_{1\le i\le d}|g_i| \le \varepsilon$), then the $l^{1,2}$ norm is equivalent to scaled l^2 norm squared:

$$\frac{1}{\left(1+\sqrt{3}c\right)\varepsilon + \sqrt{3}\tilde{c}}\|g\|_2^2 \le \|g\|_{l^{1,2}} \le \frac{1}{\sqrt{3}\tilde{c}}\|g\|_2^2.$$

On the other hand, if iterations are away from a minimizer (i.e., $\min_{1\le i\le d}|g_i| \ge L$), then the $l^{1,2}$-norm is equivalent to scaled l^1 norm:

$$\frac{1}{1+\sqrt{3}(c+\tilde{c}/L)}\|g\|_1 \le \|g\|_{l^{1,2}} \le \frac{1}{1+\sqrt{3}c}\|g\|_1.$$

These equivalences are visible in Figure 1, where we plot the level sets of $g \mapsto \|g\|_{l^{1,2}}$ at various distances from the origin. Similar mixed norm observation for signSGD was also noted by Bernstein et al. [2019]. Alternatively, under the assumptions of Lemma 2, ρ-norm can be lower bounded by a weighted l^1 norm with positive constant weights $1 - {}^{2c_i}/_{\tau} > 0$:

$$\|g\|_\rho = \sum_{i=1}^{d}(2\rho_i - 1)|g_i| \ge \sum_{i=1}^{d}(1 - {}^{2c_i}/_{\tau})|g_i|.$$

5 Convergence Theory

Now we turn to our theoretical results of sign based methods. First we give our general convergence rates under the SPB assumption. Afterwards, we extend the theory to parallel setting under the same SPB assumptions with majority vote aggregation. Finally, we explain the convergence issue of signSGD in distributed training with partitioned data and propose a new sign based method, *SSDM*, to resolve it.

Algorithm 1 SIGNSGD

```
Input: step size $\gamma_k$, current point $x_k$
$\hat{g}_k \leftarrow \mathrm{StochasticGradient}(f, x_k)$
Option 1: $x_{k+1} = x_k - \gamma_k \operatorname{sign} \hat{g}_k$
Option 2: $x_{k+1} = \arg\min\{f(x_k), f(x_k - \gamma_k \operatorname{sign} \hat{g}_k)\}$
```

Throughout the paper we assume that function $f\colon \mathbb{R}^d \to \mathbb{R}$ is nonconvex and lower bounded, i.e., $f(x) \ge f^*$ for all $x \in \mathbb{R}^d$.

5.1 Convergence analysis for $M = 1$

We start our convergence theory with single node setting, where f is smooth with some non-negative constants $(L_i)_{i=1}^d$, i.e.,

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \sum_{i=1}^d \frac{L_i}{2}(y_i - x_i)^2$$

for all x, $y \in \mathbb{R}^d$. Let $\bar{L} \coloneqq \frac{1}{d}\sum_{i=1}^d L_i$.

We now state our convergence result for signSGD (2) under the general SPB assumption.

Theorem 5 (see B.5)**.** *Under the SPB assumption, single node signSGD (Algorithm 1) with Option 1 and with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as follows*

$$\min_{0\le k<K} \mathbb{E}\|\nabla f(x_k)\|_\rho \le \frac{f(x_0) - f^*}{\gamma_0\sqrt{K}} + \frac{3\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}\,. \tag{7}$$

If $\gamma_k \equiv \gamma > 0$, we get $1/K$ convergence to a neighbourhood:

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_\rho \le \frac{f(x_0) - f^*}{\gamma K} + \frac{\gamma d\bar{L}}{2}\,. \tag{8}$$

We now comment on the above result:

• **Generalization.** Theorem 5 is the first general result on signSGD for non-convex functions without mini-batching, and with step sizes independent of the total number of iterations K. Known convergence results [Bernstein et al., 2018, 2019] on signSGD use mini-batches and/or step sizes dependent on K. Moreover, they also use unbiasedness and unimodal symmetric noise assumptions, which are stronger assumptions than our SPB assumption (see Lemma 1). Finally, Theorem 1 in [Bernstein et al., 2019] can be recovered from Theorem 5 (see Appendix C).

• **Convergence rate.** Rates (7) and (8) can be arbitrarily slow, depending on the probabilities ρ_i. This is to be expected. At one extreme, if the gradient noise was completely random, i.e., if $\rho_i \equiv 1/2$, then the ρ-norm would become identical zero for any gradient vector and rates would be trivial inequalities, leading to divergence as in the counterexample. At other extreme, if there was no gradient noise, i.e., if $\rho_i \equiv 1$, then the ρ-norm would be just the l^1 norm and we get the rate $\mathcal{O}(1/\sqrt{K})$ with respect to the l^1 norm. However, if we know that $\rho_i > 1/2$, then we can ensure that the method will eventually converge.

• **Geometry.** The presence of the ρ-norm in these rates suggests that, in general, there is no particular geometry (e.g., l^1 or l^2) associated with signSGD. Instead, the geometry is induced from the success probabilities. For example, in the case of unbiased and unimodal symmetric noise, the geometry is described by the mixture norm $l^{1,2}$.

Next, we state a general convergence rate for Algorithm 1 with Option 2.

Theorem 6 (see B.6)**.** *Under the SPB assumption, signSGD (Algorithm 1) with Option 2 and with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as follows:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_\rho \le \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \gamma_0 d\bar{L}\right].$$

In the case of $\gamma_k \equiv \gamma > 0$, the same rate as (8) is achieved.

Comparing Theorem 6 with Theorem 5, notice that one can remove the log factor from (7) and bound the average of past gradient norms instead of the minimum. On the other hand, in a big data regime, function evaluations in Algorithm 1 (Option 2, line 4) are infeasible. Clearly, Option 2 is useful *only* in the setup when one can afford function evaluations and has rough estimates about the gradients (i.e., signs of stochastic gradients). This option should be considered within the framework of derivative-free optimization.

5.2 Convergence analysis in parallel setting

In this part we present the convergence result of parallel signSGD (Algorithm 2) with majority vote introduced by Bernstein et al. [2018]. Majority vote is considered within a parameter server framework, where for each coordinate parameter server receives one sign from each node and sends back the sign sent by the majority of nodes. In parallel setting, the training data is shared among the nodes.

Algorithm 2 PARALLEL SIGNSGD WITH MAJORITY VOTE

```
**Input:** step size $\gamma_k$, current point $x_k$, # of nodes $M$
**on** each node $n$
    $\hat{g}^n(x_k) \leftarrow \text{StochasticGradient}(f, x_k)$
**on** server
    **get** sign $\hat{g}^n(x_k)$ **from** all nodes
    **send** sign $\left[\sum_{n=1}^M \operatorname{sign} \hat{g}^n(x_k)\right]$ **to** all nodes
**on** each node $n$
    $x_{k+1} = x_k - \gamma_k \operatorname{sign}\left[\sum_{n=1}^M \operatorname{sign} \hat{g}^n(x_k)\right]$
```

Known convergence results [Bernstein et al., 2018, 2019] use $\mathcal{O}(K)$ mini-batch size as well as $\mathcal{O}(1/K)$ constant step size. In the sequel we remove this limitations extending Theorem 5 to parallel training. In this case the number of nodes M get involved in geometry introducing new ρ_M-norm, which is defined by the regularized incomplete beta function I (see Appendix B.7).

Definition 2 (ρ_M-norm)**.** Let M be the number of nodes and $l := \lfloor \frac{M+1}{2} \rfloor$. Define ρ_M-norm of gradient $g(x)$ at $x \in \mathbb{R}^d$ via

$$\|g(x)\|_{\rho_M} := \sum_{i=1}^d \left(2I(\rho_i(x); l, l) - 1\right) |g_i(x)|.$$

Clearly, ρ_1-norm coincides with ρ-norm. Now we state the convergence rate of parallel signSGD with majority vote.

Theorem 7 (see B.7)**.** *Under SPB assumption, parallel signSGD (Algorithm 2) with step sizes* $\gamma_k = \gamma_0/\sqrt{k+1}$ *converges as follows*

$$\min_{0 \le k < K} \mathbb{E}\|\nabla f(x_k)\|_{\rho_M} \le \frac{f(x_0) - f^*}{\gamma_0 \sqrt{K}} + \frac{3\gamma_0 d\bar{L}}{2} \frac{\log K}{\sqrt{K}}. \tag{9}$$

For constant step sizes $\gamma_k \equiv \gamma > 0$*, we have convergence up to a level proportional to step size* γ*:*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_{\rho_M} \le \frac{f(x_0) - f^*}{\gamma K} + \frac{\gamma d\bar{L}}{2}. \tag{10}$$

• **Speedup with respect to M.** Note that, in parallel setting with M nodes, the only difference in convergence rates (9) and (10) is the modified ρ_M-norm measuring the size of gradients. Using Hoeffding's inequality, we show (see Appendix B.8) that $\|g(x)\|_{\rho_M} \to \|g(x)\|_1$ exponentially fast as $M \to \infty$, namely

$$\left(1 - e^{-(2\rho(x)-1)^2 l}\right) \|g(x)\|_1 \le \|g(x)\|_{\rho_M} \le \|g(x)\|_1,$$

where $\rho(x) = \min_{1\le i\le d} \rho_i(x) > 1/2$. To appreciate the speedup with respect to M, consider the noise-free case as a baseline, for which $\rho_i \equiv 1$ and $\|g(x)\|_{\rho_M} \equiv \|g(x)\|_1$. Then, the above inequality implies that M parallel machines reduce the variance of gradient noise exponentially fast.

• **Number of Nodes.** Theoretically there is no difference between $2l-1$ and $2l$ nodes, and this is not a limitation of the analysis. Indeed, as it is shown in the proof, expected sign vector at the master with $M = 2l-1$ nodes is the same as with $M = 2l$ nodes:

$$\mathbb{E}\,\mathrm{sign}(\hat{g}_i^{(2l)} \cdot g_i) = \mathbb{E}\,\mathrm{sign}(\hat{g}_i^{(2l-1)} \cdot g_i),$$

where $\hat{g}^{(M)}$ is the sum of stochastic sign vectors aggregated from nodes. Intuitively, majority vote with even number of nodes, e.g. $M = 2l$, fails to provide any sign with little probability (it is the probability of half nodes voting for $+1$, and half nodes voting for -1). However, if we remove one node, e.g. $M = 2l-1$, then master receives one sign-vote less but gets rid of that little probability of failing the vote (sum of odd number of ±1 cannot vanish).

5.3 Distributed training with partitioned data

First, we briefly discuss the fundamental issue of signSGD in distributed environment and then present our new sign based method which resolves that issue.

5.3.1 The issue with distributed signSGD

Consider distributed training where each machine $n \in \{1, 2, \dots, M\}$ has its own loss function $f_n(x)$. We argue that in this setting even signGD (with full-batch gradients and no noise) can fail to converge. Indeed, let us multiply each loss function $f_n(x)$ of nth node by an arbitrary positive scalars $w_n > 0$. Then the landscape (in particular, stationary points) of the overall loss function

$$f^w(x) \coloneqq \tfrac{1}{M} \sum_{n=1}^{M} w_n f_n(x)$$

can change arbitrarily while the iterates of signGD remain the same as the master server aggregates the same signs $\mathrm{sign}(w_n \nabla f_n(x)) = \mathrm{sign}\, \nabla f_n(x)$ regardless of the scalars $w_n > 0$. Thus, distributed signGD is unable to sense the weights $w_n > 0$ modifying total loss function f^w and cannot guarantee approximate stationary point unless loss functions f_n have some special structures.

5.3.2 Novel sign-based method for distributed training

The above issue of distributed signSGD stems from the biasedness of the sign operator which completely ignores the magnitudes of local gradients of all nodes. We resolve this issue by designing a novel distributed sign-based method–*Stochastic Sign Descent with Momentum (SSDM)*–including two additional layers: *stochastic sign* and *momentum.*

Motivated by SPB assumption, we introduce our new notion of *stochastic sign* to replace the usual deterministic sign.

Definition 3 (Stochastic Sign)**.** We define the stochastic sign operator $\widetilde{\operatorname{sign}} : \mathbb{R}^d \to \mathbb{R}^d$ via

$$\left(\widetilde{\operatorname{sign}}\, g\right)_i = \begin{cases} +1, & \text{with probability } \frac{1}{2} + \frac{1}{2}\frac{g_i}{\|g\|} \\ -1, & \text{with probability } \frac{1}{2} - \frac{1}{2}\frac{g_i}{\|g\|} \end{cases}$$

for $1 \le i \le d$ and $\widetilde{\operatorname{sign}}\,\mathbf{0} = \mathbf{0}$ with probability 1.

Technical importance of stochastic $\widetilde{\operatorname{sign}}$ is twofold. First, it satisfies the SPB assumption automatically, that is

$$\operatorname{Prob}((\widetilde{\operatorname{sign}}\, g)_i = \operatorname{sign} g_i) = \frac{1}{2} + \frac{1}{2}\frac{|g_i|}{\|g\|} > \frac{1}{2},$$

if $g_i \neq 0$. Second, unlike the deterministic sign operator, it is unbiased with scaling factor $\|g\|$, namely $\mathbb{E}[\|g\|\, \widetilde{\operatorname{sign}}\, g] = g$. We describe our SSDM method formally in Algorithm 3.

Algorithm 3 Stochastic Sign Descent with Momentum (SSDM)

```
Input: step size parameter $\gamma$, momentum parameter $\beta$, # of nodes $M$
Initialize: $x_0 \in \mathbb{R}^d$, $m_{-1}^n = \hat{g}_0^n$ for all $n \in \{1, 2, \dots, M\}$
for $k = 0, 1, \dots, K-1$ do
    on each node $n$
        $\hat{g}_k^n \leftarrow \mathrm{StochasticGradient}(f_n, x_k)$        Local sub-sampling
        $m_k^n = \beta m_{k-1}^n + (1-\beta)\hat{g}_k^n$        Update the momentum
        send $s_k^n := \widetilde{\operatorname{sign}}\, m_k^n$ to the server        Communicate to the server
    on server
        send $s_k := \sum_{n=1}^M s_k^n$ to all nodes        Communicate to all nodes
    on each node $n$
        $x_{k+1} = x_k - \frac{\gamma}{M} s_k$        Main step: Update the global model
end for
```

Consider the optimization problem (1), where each node n owns only the data associated with loss function $f_n \colon \mathbb{R}^d \to \mathbb{R}$, which is non-convex and L^n-smooth. We model stochastic gradient oracle using the standard bounded variance condition defined below:

Assumption 2 (Bounded Variance)**.** For any $x \in \mathbb{R}^d$, each node n has access to an unbiased estimator $\hat{g}^n(x)$ with bounded variance $(\sigma^n)^2 \ge 0$, namely

$$\mathbb{E}\left[\hat{g}^n(x)\right] = \nabla f_n(x), \quad \mathbb{E}\left[\|\hat{g}^n(x) - \nabla f_n(x)\|^2\right] \le (\sigma^n)^2.$$

Now, we present our convergence result for SSDM method.

Theorem 8 (see B.9)**.** *Under Assumption 2, $K \ge 1$ iterations of SSDM (Algorithm 3) with momentum parameter $\beta = 1 - \frac{1}{\sqrt{K}}$ and step-size $\gamma = \frac{1}{K^{3/4}}$ guarantee*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x^k)\| \le \frac{1}{K^{1/4}}\left[3\delta_f + 16\tilde{\sigma} + 8\tilde{L}\sqrt{d} + \frac{3\tilde{L}d}{\sqrt{K}}\right],$$

where $\delta_f = f(x_0) - f^*, \tilde{\sigma} = \frac{1}{M}\sum\limits_{n=1}^{M} \sigma^n, \tilde{L} = \frac{1}{M}\sum\limits_{n=1}^{M} L^n$.

• **Optimal rate using sign bits only.** Note that, for non-convex distributed training, SSDM has the same optimal asymptotic rate $\mathcal{O}(\varepsilon^{-4})$ as SGD. In contrast, signSGD and its momentum version Signum Bernstein et al. [2018, 2019] were not analyzed in distributed setting where data is partitioned between nodes and require increasingly larger mini-batches over the course of training. A general approach to handle biased compression operators, satisfying certain contraction property, is the *error feedback (EF)* mechanism proposed by Seide et al. [2014]. In particular, EF-signSGD method of Karimireddy et al. [2019] fixes the convergence issues of signSGD in single node setup, overcoming SBP assumption. Furthermore, for distributed training, Tang et al. [2019] applied the error feedback trick both for the server and nodes in their DoubleSqueeze method maintaining the same asymptotic rate with bi-directional compression. However, in these methods, the contraction property of compression operator used by error feedback forces to communicate the magnitudes of local stochastic gradients together with the signs. This is not the case for sign-based methods considered in this work, where only sign bits are communicated between nodes and server.

• **Noisy signSGD.** In some sense, stochastic sign operator (see Definition 3) can be viewed as noisy version of standard deterministic sign operator and, similarly, our SSDM method can be viewed as noisy variant of signSGD with momentum. This observation reveals a connection to the noisy signSGD method of Chen et al. [2020]. Despite some similarities between the two methods, there are several technical aspects that SSDM excels their noisy signSGD. First, the noise they add is *artificial* and requires a special care: too much noise blows the convergence, too little noise is unable to shrink the gap between median and mean. Moreover, as it is discussed in their paper, the variance of the noise must depend on K (total number of iterations) and tend to ∞ with K to guarantee convergence to stationary points in the limit. Meanwhile, the noise of SSDM is *natural* and does not need to be adjusted. Next, the convergence bound (17) of [Chen et al., 2020] is harder to interpret than the bound in our Theorem 4 involving l^2 norms of the gradients *only*. Besides, the convergence rate with respect to squared l^2 norm is $\mathcal{O}(d^{3/4}/K^{1/4})$, while the rate of SSDM with respect to squared l^2 norm is $\mathcal{O}(d/\sqrt{K})$, which is $\mathcal{O}(K^{1/4}/d^{1/4})$ times *faster*. Lastly, it is explicitly written before Theorem 5 that the analysis assumes *full* gradient computation for all nodes. In contrast, SSDM is analyzed under a more general stochastic gradient oracle.

• **All-reduce compatible.** In contrast to signSGD with majority vote aggregation, SSDM supports partial aggregation of compressed stochastic signs s_k^n. In other words, compressed signs s_k^n can be directly summed without additional decompression-compression steps. This allows SSDM to be implemented with efficient *all-reduce* operation instead of slower *all-gather* operation. Besides SSDM, only a few compression schemes in the literature satisfy this property and can be implemented with *all-reduce* operation, e.g., SGD with random sparsification Wangni et al. [2018], GradiVeQ Yu et al. [2018], PowerSGD Vogels et al. [2019].

Finally, we show that the improved convergence theory and low communication cost of SSDM is due to the use of *both* stochastic sign operator and momentum term.

• **SSDM without stochastic sign.** If we replace stochastic sign by deterministic sign in SSDM, then the resulting method *can provably diverge* even when full gradients are computed by all nodes. In fact, the counterexample (5) in Section 3.2 can be easily extended to distributed setting and can handle momentum. Indeed, consider $M = 2$ nodes owning functions $f_n(x) = \langle a_n, x\rangle^2$, $n = 1, 2$ with a_1, a_2 as defined in (5) and initial point $x_0 \in Z = \{(z_1, z_2) \colon z_1 + z_2 = 2\}$. Since $\nabla f_n(x) = 2\langle a_n, x\rangle a_n \in span(a_n)$, we imply $m_k^n \in span(a_n)$ for any value of parameter β and for all iterate $k \geq 0$ (see lines 2 and 6 of Algorithm 3). Hence, $\operatorname{sign} m_k^n = \pm \operatorname{sign} a_n = \pm \left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right]$. Since $s_k = \operatorname{sign} m_k^1 + \operatorname{sign} m_k^2 \in span(\left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right])$ (see line 9), this means that the method is again stuck along the line Z as $\frac{\gamma}{M} s_k \in span(\left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right])$ (see line 11) for any value of γ.

• **SSDM without momentum.** It is possible to obtain the same asymptotic convergence rate without the momentum term (i.e., $\beta = 0$). In this case, if all nodes also send the norms $\|\hat{g}_k^n\|$ to the server then the method can be analyzed by a standard analysis of distributed SGD with an unbiased compression. However, the drawback of this approach is the *higher communication cost.* While the overhead of worker-to-server communication is negligible (one extra float), the reverse server-to-worker communication becomes costly as the aggregated updates are dense (all entries are floats) as opposed to the original SSDM (all entries are integers).

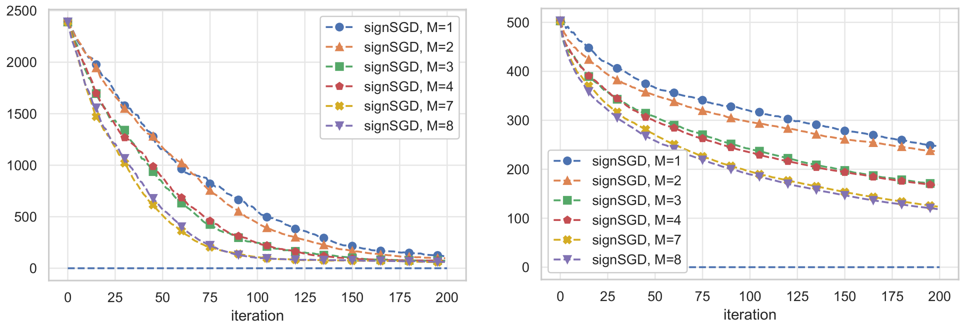

Figure 2: Experiments on distributed signSGD with majority vote using Rosenbrock function. Plots show function values with respect to iterations averaged over 10 repetitions. Left plot used constant step size $\gamma = 0.02$, right plot used variable step size with $\gamma_0 = 0.02$. We set mini-batch size 1 and used the same initial point. Dashed blue lines mark the minimum.

6 Experiments

We verify several aspects of our theoretical results experimentally using the MNIST dataset with feed-forward neural network (FNN) and the well known Rosenbrock (non-convex) function with $d = 10$ variables:

$$f(x) = \sum_{i=1}^{d-1} f_i(x), \quad \text{where} \quad f_i(x) = 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2.$$

6.1 Minimizing the Rosenbrock function

The Rosenbrock function is a classic example of non-convex function, which is used to test the performance of optimization methods. We chose this low dimensional function in order to estimate the success probabilities effectively in a reasonable time and to expose theoretical connection.

Stochastic formulation of the minimization problem for Rosenbrock function is as follows: at any point $x \in \mathbb{R}^d$ we have access to *biased* stochastic gradient $\hat{g}(x) = \nabla f_i(x) + \xi$, where index i is chosen uniformly at random from $\{1, 2, \ldots, d-1\}$ and $\xi \sim \mathcal{N}(0, \nu^2 I)$ with $\nu > 0$.

Figure 2 illustrates the effect of multiple nodes in distributed training with majority vote. As we see increasing the number of nodes improves the convergence rate. It also supports the claim that in expectation there is no improvement from $2l - 1$ nodes to $2l$ nodes.

Figure 3 shows the robustness of SPB assumption in the convergence rate (8) with constant step size. We exploited four levels of noise in each column to demonstrate the correlation between success probabilities and convergence rate. In the first experiment (first column) SPB assumption is violated strongly and the corresponding rate shows divergence. In the second column, probabilities still violating SPB assumption are close to the threshold and the rate shows oscillations. Next columns express the improvement in rates when success probabilities are pushed to be close to 1. More experiments on the Rosenbrock function are moved to Appendix A.

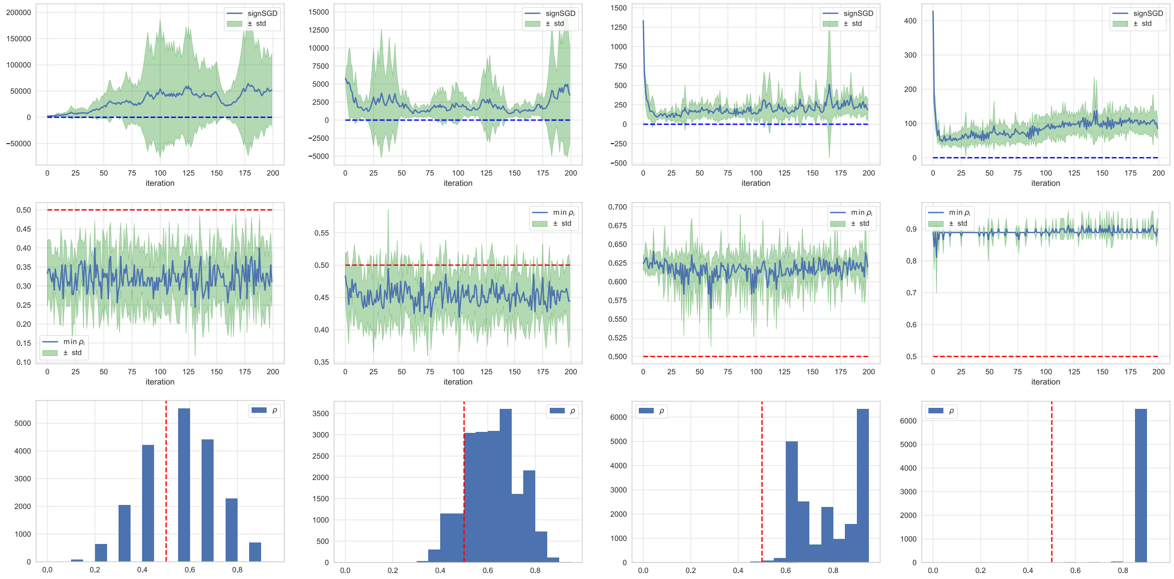

Figure 3: Performance of signSGD with constant step size ($\gamma = 0.25$) under four different noise levels (mini-batch size 1, 2, 5, 8) using Rosenbrock function. Each column represent a separate experiment with function values, evolution of minimum success probabilities and the histogram of success probabilities throughout the iteration process. Dashed blue line in the first row is the minimum value. Dashed red lines in second and third rows are thresholds 1/2 of success probabilities. The shaded area in first and second rows shows standard deviation obtained from ten repetitions.

6.2 Training FNN on the MNIST dataset

We trained a single layer feed-forward network on the MNIST with two different batch construction strategies. The first construction is the standard way of training networks: before each epoch we shuffle the training dataset and choose batches sequentially. In the second construction, first we split the training dataset into two parts, images with labels 0, 1, 2, 3, 4 and images with labels 5, 6, 7, 8, 9. Then each batch of images were chosen from one of these parts with equal probabilities. We make the following observations based on our experiments depicted in Figure 4 and Figure 5.

• **Convergence with multi-modal and skewed gradient distributions.** Due to the split batch construction strategy we unfold multi-modal and asymmetric distributions for stochastic gradients in Figure 4. With this experiment we conclude that sign based methods can converge under various gradient distributions which is allowed from our theory.

• **Effectiveness in the early stage of training.** Both experiments show that in the beginning of the training, signSGD is more efficient than SGD when we compare accuracy against communication. This observation is supported by the theory as at the start of the training success probabilities are bigger (see Lemma 1) and lower bound for mini-batch size is smaller (see Lemma 3).

• **Bigger batch size, better convergence.** Figure 5 shows that the training with larger batch size improves the convergence as backed by the theory (see Lemmas 2 and 3).

• **Generalization effect.** Another aspect of sign based methods which has been observed to be problematic, in contrast to SGD, is the generalization ability of the model (see also [Balles and Hennig, 2018], Section 6.2 Results). In the experiment with standard batch construction (see Figure 5) we notice that test accuracy is growing with training accuracy. However, in the other experiment with split batch construction (see Figure 4), we found that test accuracy does not get improved during the second half of the training while train accuracy grows consistently with slow pace.

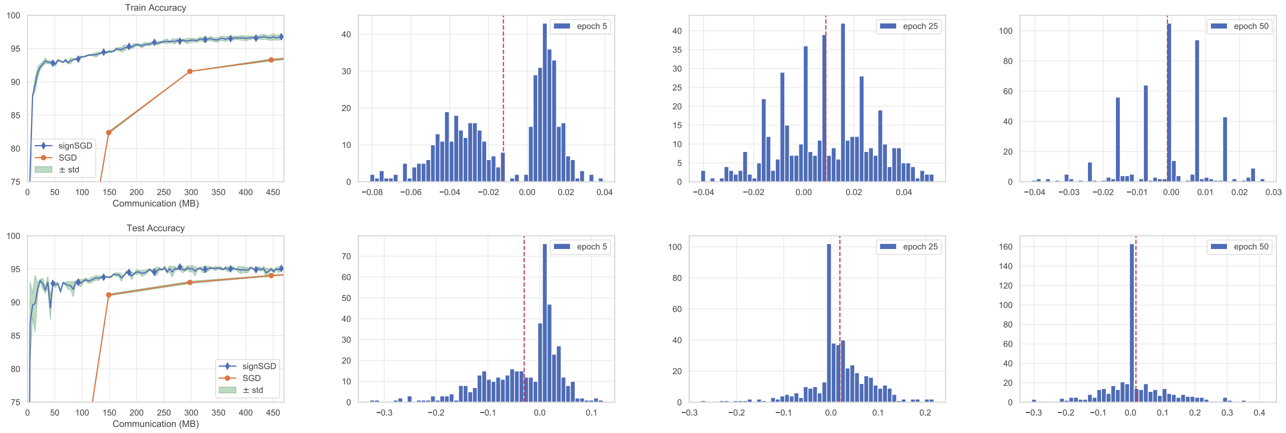

Figure 4: Convergence of signSGD and comparison with SGD on the MNIST dataset using the split batch construction strategy. The budget of gradient communication (MB) is fixed and the network is a single hidden layer FNN. We first tuned the constant step size over logarithmic scale $\{1, 0.1, 0.01, 0.001, 0.0001\}$ and then fine tuned it. First column shows train and test accuracies with mini-batch size 128 and averaged over 3 repetitions. We chose two weights (empirically, most of the network biases would work) and plotted histograms of stochastic gradients before epochs 5, 25 and 50. Dashed red lines on histograms indicate the average values.

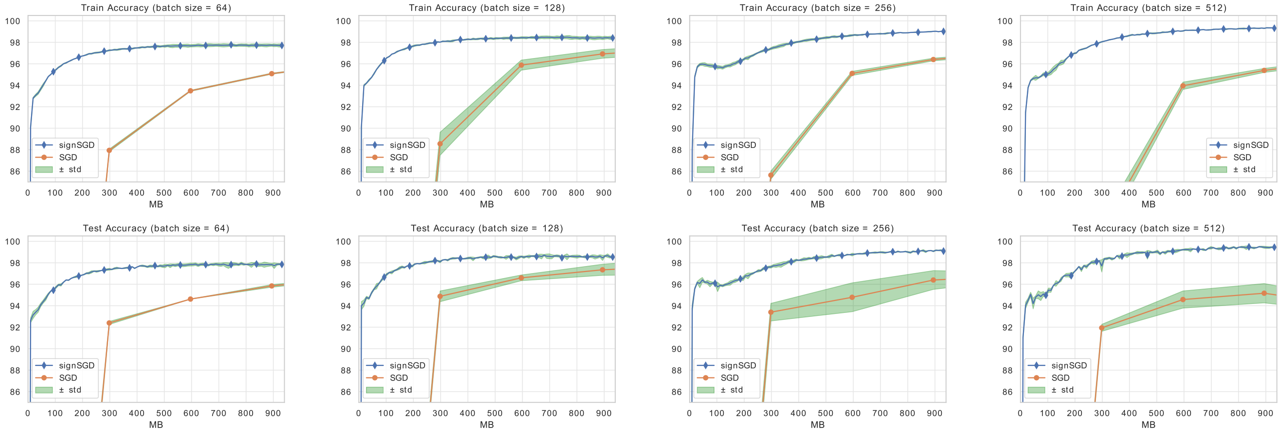

Figure 5: Comparison of signSGD and SGD on the MNIST dataset with a fixed budget of gradient communication (MB) using single hidden layer FNN and the standard batch construction strategy. For each batch size, we first tune the constant step size over logarithmic scale $\{10, 1, 0.1, 0.01, 0.001\}$ and then fine tune it. Shaded area shows the standard deviation from 3 repetition.

Acknowledgments

Mher Safaryan thanks El Houcine Bergou, Konstantin Mishchenko and Samuel Horváth for helpful discussions about the topic and their feedback on the paper.

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 30*, pages 1709–1720, 2017.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, pages 404–413, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 560–569. PMLR, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Léon Bottou and Yann Le Cun. Large scale online learning. In *Advances in Neural Information Processing Systems*, 2003.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 111–119, 2015.

Xiangyi Chen, Tiancong Chen, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. Distributed training with heterogeneous data: Bridging median- and mean-based algorithms. In *34th Conference on Neural Information Processing Systems*, 2020.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268, Virtual, 13–18 Jul 2020. PMLR. URL `http://proceedings.mlr.press/v119/cutkosky20b.html`.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Journal of Machine Learning Research*, page 2121–2159, 2011.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. In *SIAM Journal on Optimization*, volume 23(4), page 2341–2368, 2013.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3252–3261, 2019.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. In *arXiv preprint arXiv:1806.06573*, 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 1097–1105, 2012.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. In *arXiv preprint arXiv:1901.09269*, 2019.

Xun Qian, Peter Richtárik, Robert Mansel Gower, Alibek Sailanbayev, Nicolas Loizou, and Egor Shulgin. SGD with arbitrary sampling: General analysis and improved rates. In *International Conference on Machine Learning*, 2019.

Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2019.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The Rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

Herbert Robbins and Sutton Monro. A stochastic approximation method. In *The Annals of Mathematical Statistics*, volume 22(3), pages 400–407, 1951.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. In *Neural networks*, volume 61, page 85–117, 2015.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. In *arXiv preprint arXiv:1308.6370*, 2013.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. In *Mathematical Programming*, volume 162(1-2), page 83–112, 2017.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Irina Shevtsova. On the absolute constants in the berry–esseen type inequalities for identically distributed summands. In *arXiv preprint arXiv:1111.6554*, 2011.

Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. `DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Int. Conf. Machine Learning*, volume PMLR 97, pages 6155–6165, 2019.

Tijmen Tieleman and Geoffrey E. Hinton. RMSprop. In *Coursera: Neural Networks for Machine Learning, Lecture 6.5*, 2012.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models (and an accelerated perceptron). In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR*, volume 89, 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *33th Advances in Neural Information Processing Systems*, 2019.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *32th Advances in Neural Information Processing Systems*, 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*, page 1509–1519, 2017.

Ashia Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annavaram, and Salman Avestimehr. GradiVeQ: Vector quantization for bandwidth-efficient gradient aggregation in distributed CNN training. In *32th Advances in Neural Information Processing Systems*, 2018.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 9815–9825, 2018.

Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. In *arXiv e-prints, arXiv:1212.5701*, 2012.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, page 4035–4043, 2017.

Appendix

A Additional Experiments

In this section we present several more experiments on the Rosenbrock function for further insights.

Figure 6 experiments with the same setup but variable learning rate. In Figure 7, we investigated the size of the neighborhood with respect to step size.

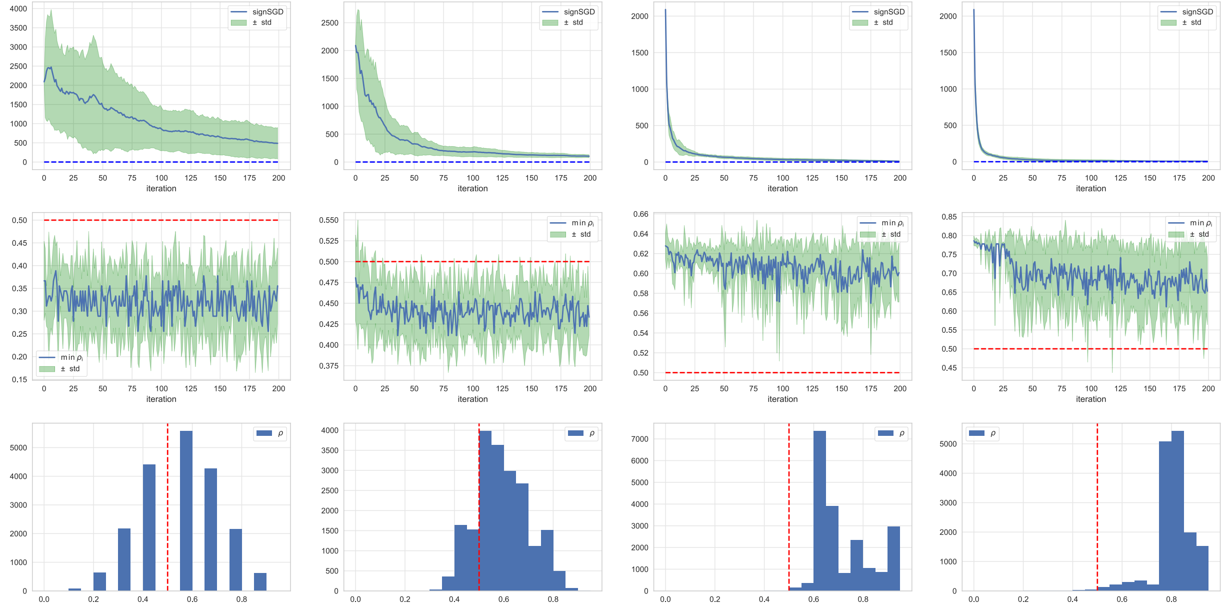

Figure 6: Performance of signSGD with variable step size ($\gamma_0 = 0.25$) under four different noise levels (mini-batch size 1, 2, 5, 7) using Rosenbrock function. As in the experiments of Figure 3 with constant step size, these plots show the relationship between success probabilities and the convergence rate (7). In low success probability regime (first and second columns) we observe oscillations, while in high success probability regime (third and forth columns) oscillations are mitigated substantially.

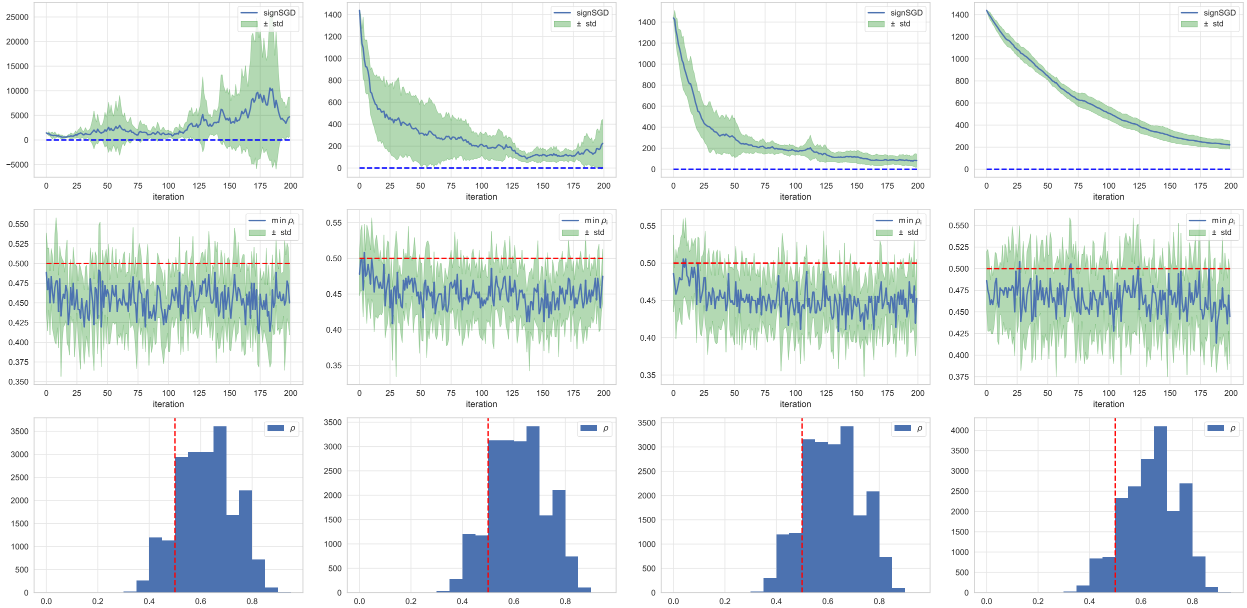

Figure 7: In this part of experiments we investigated convergence rate (8) to a neighborhood of the solution. We fixed gradient noise level by setting mini-batch size 2 and altered the constant step size. For the first column we set bigger step size $\gamma = 0.25$ to detect the divergence (as we slightly violated SPB assumption). Then for the 2nd and 3rd columns we set $\gamma = 0.1$ and $\gamma = 0.05$ to expose the convergence to a neighborhood. For the forth column we set even smaller step size $\gamma = 0.01$ to observe a slower convergence.

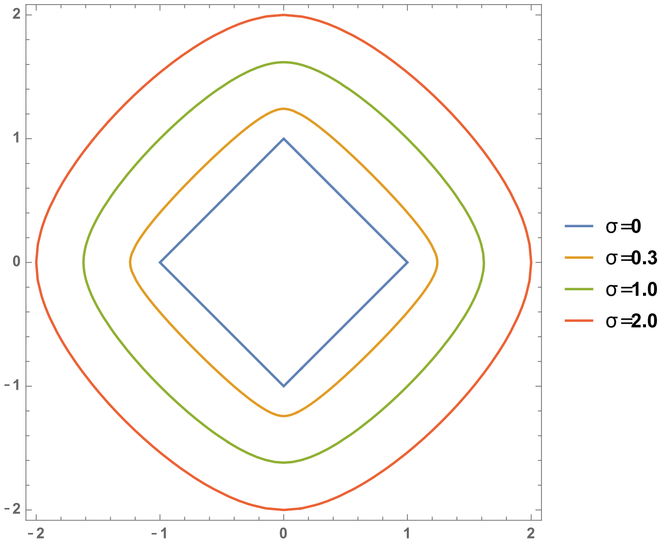

Figure 8: Unit balls in $l^{1,2}$ norm (6) with different noise levels.

B Proofs

B.1 Sufficient conditions for SPB: Proof of Lemma 1

Here we state the well-known Gauss's inequality on unimodal distributions[2].

Theorem 9 (Gauss's inequality)**.** *Let X be a unimodal random variable with mode m, and let σ_m^2 be the expected value of $(X-m)^2$. Then for any positive value of r,*

$$\operatorname{Prob}(|X-m|>r) \le \begin{cases} \frac{4}{9}\left(\frac{\sigma_m}{r}\right)^2, & \textit{if } r \ge \frac{2}{\sqrt{3}}\sigma_m \\ 1-\frac{1}{\sqrt{3}}\frac{r}{\sigma_m}, & \textit{otherwise} \end{cases}$$

Applying this inequality on unimodal and symmetric distributions, direct algebraic manipulations give the following bound:

$$\operatorname{Prob}(|X-\mu|\le r) \ge \begin{cases} 1-\frac{4}{9}\left(\frac{\sigma}{r}\right)^2, & \text{if } \frac{\sigma}{r} \le \frac{\sqrt{3}}{2} \\ \frac{1}{\sqrt{3}}\frac{r}{\sigma}, & \text{otherwise} \end{cases} \ge \frac{r/\sigma}{r/\sigma+\sqrt{3}},$$

where $m=\mu$ and $\sigma_m^2=\sigma^2$ are the mean and variance of unimodal, symmetric random variable X, and $r\ge 0$. Now, using the assumption that each $\hat{g}_i(x)$ has unimodal and symmetric distribution, we apply this bound for $X=\hat{g}_i(x)$, $\mu=g_i(x), \sigma^2=\sigma_i^2(x)$ and get a bound for success probabilities

$$\begin{aligned}
\operatorname{Prob}(\operatorname{sign}\hat{g}_i = \operatorname{sign} g_i) &= \begin{cases} \operatorname{Prob}(\hat{g}_i \ge 0), & \text{if } g_i>0 \\ \operatorname{Prob}(\hat{g}_i \le 0), & \text{if } g_i<0 \end{cases} \\
&= \begin{cases} \frac{1}{2}+\operatorname{Prob}(0\le \hat{g}_i \le g_i), & \text{if } g_i>0 \\ \frac{1}{2}+\operatorname{Prob}(g_i\le \hat{g}_i \le 0), & \text{if } g_i<0 \end{cases} \\
&= \begin{cases} \frac{1}{2}+\frac{1}{2}\operatorname{Prob}(0\le \hat{g}_i \le 2g_i), & \text{if } g_i>0 \\ \frac{1}{2}+\frac{1}{2}\operatorname{Prob}(2g_i\le \hat{g}_i \le 0), & \text{if } g_i<0 \end{cases} \\
&= \frac{1}{2}+\frac{1}{2}\operatorname{Prob}(|\hat{g}_i-g_i|\le |g_i|) \\
&\ge \frac{1}{2}+\frac{1}{2}\frac{|g_i|/\sigma_i}{|g_i|/\sigma_i+\sqrt{3}} \\
&= \frac{1}{2}+\frac{1}{2}\frac{|g_i|}{|g_i|+\sqrt{3}\sigma_i}
\end{aligned}$$

Improvement on Lemma 1 and $l^{1,2}$ norm: The bound after Gauss inequality can be improved including a second order term

$$\operatorname{Prob}(|X-\mu|\le r) \ge \begin{cases} 1-\frac{4}{9}\left(\frac{\sigma}{r}\right)^2, & \text{if } \frac{\sigma}{r} \le \frac{\sqrt{3}}{2} \\ \frac{1}{\sqrt{3}}\frac{r}{\sigma}, & \text{otherwise} \end{cases} \ge 1-\frac{1}{1+{}^r\!/\!\sqrt{3}\sigma+({}^r\!/\!\sqrt{3}\sigma)^2}.$$

Indeed, letting $z := {}^r\!/\!\sqrt{3}\sigma \ge {}^2\!/\!3$, we get $1-\frac{4}{9}\frac{1}{3z^2} \ge 1-\frac{1}{1+z+z^2}$ as it reduces to $23z^2-4z-4\ge 0$. Otherwise, if $0\le z\le {}^2\!/\!3$, then $z \ge 1-\frac{1}{1+z+z^2}$ as it reduces to $1\ge 1-z^3$. The improvement is tighter as

$$\frac{r/\sigma}{r/\sigma+\sqrt{3}} = 1-\frac{1}{1+{}^r\!/\!\sqrt{3}\sigma} \le 1-\frac{1}{1+{}^r\!/\!\sqrt{3}\sigma+({}^r\!/\!\sqrt{3}\sigma)^2}.$$

[2] see `https://en.wikipedia.org/wiki/Gauss%27s_inequality`

Hence, continuing the proof of Lemma 1, we get

$$\operatorname{Prob}(\operatorname{sign}\hat{g}_i = \operatorname{sign} g_i) \geq 1 - \frac{1}{2}\frac{1}{1 + {}^{|g_i|}/_{\sqrt{3}\sigma_i} + ({}^{|g_i|}/_{\sqrt{3}\sigma_i})^2}$$

and we could have defined $l^{1,2}$-norm in a bit more complicated form as

$$\|g\|_{l^{1,2}} := \sum_{i=1}^{d}\left(1 - \frac{1}{1 + {}^{|g_i|}/_{\sqrt{3}\sigma_i} + ({}^{|g_i|}/_{\sqrt{3}\sigma_i})^2}\right)|g_i|.$$

B.2 Sufficient conditions for SPB: Proof of Lemma 2

Let $\hat{g}^{(\tau)}$ be the gradient estimator with mini-batch size τ. It is known that the variance for $\hat{g}^{(\tau)}$ is dropped by at least a factor of τ, i.e.

$$\mathbb{E}[(\hat{g}_i^{(\tau)} - g_i)^2] \leq \frac{\sigma_i^2}{\tau}.$$

Hence, estimating the failure probabilities of $\operatorname{sign}\hat{g}^{(\tau)}$ when $g_i \neq 0$, we have

$$\begin{aligned}\operatorname{Prob}(\operatorname{sign}\hat{g}_i^{(\tau)} \neq \operatorname{sign} g_i) &= \operatorname{Prob}(|\hat{g}_i^{(\tau)} - g_i| = |\hat{g}_i^{(\tau)}| + |g_i|) \\ &\leq \operatorname{Prob}(|\hat{g}_i^{(\tau)} - g_i| \geq |g_i|) \\ &= \operatorname{Prob}((\hat{g}_i^{(\tau)} - g_i)^2 \geq g_i^2) \\ &\leq \frac{\mathbb{E}[(\hat{g}_i^{(\tau)} - g_i)^2]}{g_i^2} \\ &= \frac{\sigma_i^2}{\tau g_i^2},\end{aligned}$$

which imples

$$\rho_i = \operatorname{Prob}(\operatorname{sign}\hat{g}_i = \operatorname{sign} g_i) \geq 1 - \frac{\sigma_i^2}{\tau g_i^2} \geq 1 - \frac{c_i}{\tau}.$$

B.3 Sufficient conditions for SPB: Proof of Lemma 3

The proof of this lemma is the most technical one. We will split the derivation into three lemmas providing some intuition on the way. The first two lemmas establish success probability bounds in terms of mini-batch size. Essentially, we present two methods: one works well in the case of small randomness, while the other one in the case of non-small randomness. In the third lemma, we combine those two bounds to get the condition on mini-batch size ensuring SPB assumption.

Lemma 10. *Let $X_1, X_2, \ldots, X_\tau$ be i.i.d. random variables with non-zero mean $\mu := \mathbb{E}X_1 \neq 0$, finite variance $\sigma^2 := \mathbb{E}|X_1 - \mu|^2 < \infty$. Then for any mini-batch size $\tau \geq 1$*

$$\operatorname{Prob}\left(\operatorname{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \operatorname{sign}\mu\right) \geq 1 - \frac{\sigma^2}{\tau\mu^2}. \tag{11}$$

Proof. Without loss of generality, we assume $\mu > 0$. Then, after some adjustments, the proof follows from the Chebyshev's inequality:

$$\begin{aligned}
\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) &= \mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0\right) \\
&\geq \mathrm{Prob}\left(\left|\frac{1}{\tau}\sum_{i=1}^{\tau} X_i - \mu\right| < \mu\right) \\
&= 1 - \mathrm{Prob}\left(\left|\frac{1}{\tau}\sum_{i=1}^{\tau} X_i - \mu\right| \geq \mu\right) \\
&\geq 1 - \frac{1}{\mu^2}\mathrm{Var}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] \\
&= 1 - \frac{\sigma^2}{\tau\mu^2},
\end{aligned}$$

where in the last step we used independence of random variables $X_1, X_2, \dots, X_\tau$. □

Obviously, bound (11) is not optimal for big variance as it becomes a trivial inequality. In the case of non-small randomness a better bound is achievable additionally assuming the finiteness of 3th central moment.

Lemma 11. *Let $X_1, X_2, \dots, X_\tau$ be i.i.d. random variables with non-zero mean $\mu := \mathbb{E}X_1 \neq 0$, positive variance $\sigma^2 := \mathbb{E}|X_1 - \mu|^2 > 0$ and finite 3th central moment $\nu^3 := \mathbb{E}|X_1 - \mu|^3 < \infty$. Then for any mini-batch size $\tau \geq 1$*

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) \geq \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{|\mu|\sqrt{\tau}}{\sqrt{2}\sigma}\right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right), \tag{12}$$

where error function erf *is defined as*

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt, \quad x \in \mathbb{R}.$$

Proof. Again, without loss of generality, we may assume that $\mu > 0$. Informally, the proof goes as follows. As we have an average of i.i.d. random variables, we approximate it (in the sense of distribution) by normal distribution using the Central Limit Theorem (CLT). Then we compute success probabilities for normal distribution with the error function erf. Finally, we take into account the approximation error in CLT, from which the third term with negative sign appears. More formally, we apply Berry–Esseen inequality[3] on the rate of approximation in CLT [Shevtsova, 2011]:

$$\left|\mathrm{Prob}\left(\frac{1}{\sigma\sqrt{\tau}}\sum_{i=1}^{\tau}(X_i - \mu) > t\right) - \mathrm{Prob}\left(N > t\right)\right| \leq \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}}, \quad t \in \mathbb{R},$$

where $N \sim \mathcal{N}(0,1)$ has the standard normal distribution. Setting $t = -\mu\sqrt{\tau}/\sigma$, we get

$$\left|\mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0\right) - \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma}\right)\right| \leq \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}}. \tag{13}$$

[3] see `https://en.wikipedia.org/wiki/Berry-Esseen_theorem`

It remains to compute the second probability using the cumulative distribution function of normal distribuition and express it in terms of the error function:

$$\begin{aligned}\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) &= \mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0\right) \\ &\overset{(13)}{\geq} \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma}\right) - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\mu\sqrt{\tau}/\sigma}^{\infty} e^{-t^2/2}\,dt - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\ &= \frac{1}{2}\left(1 + \sqrt{\frac{2}{\pi}}\int_0^{\mu\sqrt{\tau}/\sigma} e^{-t^2/2}\,dt - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right) \\ &= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma}\right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right).\end{aligned}$$

□

Clearly, bound (12) is better than (11) when randomness is high. On the other hand, bound (12) is not optimal for small randomness ($\sigma \approx 0$). Indeed, one can show that in a small randomness regime, while both variance σ^2 and third moment ν^3 are small, the ration ν/σ might blow up to infinity producing trivial inequality. For instance, taking $X_i \sim \mathrm{Bernoulli}(p)$ and letting $p \to 1$ gives $\nu/\sigma = O\left((1-p)^{-1/6}\right)$. This behaviour stems from the fact that we are using CLT: less randomness implies slower rate of approximation in CLT.

As a result of these two bounds on success probabilities, we conclude a condition on mini-batch size for the SPB assumption to hold.

Lemma 12. *Let $X_1, X_2, \ldots, X_\tau$ be i.i.d. random variables with non-zero mean $\mu \neq 0$ and finite variance $\sigma^2 < \infty$. Then*

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) > \frac{1}{2}, \quad \text{if} \quad \tau > 2\min\left(\frac{\sigma^2}{\mu^2}, \frac{\nu^3}{|\mu|\sigma^2}\right), \tag{14}$$

where ν^3 is (possibly infinite) 3th central moment.

Proof. First, if $\sigma = 0$ then the lemma holds trivially. If $\nu = \infty$, then it follows immediately from Lemma 10. Assume both σ and ν are positive and finite.

In case of $\tau > 2\sigma^2/\mu^2$ we apply Lemma 10 again. Consider the case $\tau \leq 2\sigma^2/\mu^2$, which implies $\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma} \leq 1$. It is easy to check that $\mathrm{erf}(x)$ is concave on $[0,1]$ (in fact on $[0,\infty)$), therefore $\mathrm{erf}(x) \geq \mathrm{erf}(1)x$ for any $x \in [0,1]$. Setting $x = \frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma}$ we get

$$\mathrm{erf}\left(\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma}\right) \geq \frac{\mathrm{erf}(1)}{\sqrt{2}}\frac{\mu\sqrt{\tau}}{\sigma},$$

which together with (12) gives

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) \geq \frac{1}{2}\left(1 + \frac{\mathrm{erf}(1)}{\sqrt{2}}\frac{\mu\sqrt{\tau}}{\sigma} - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right).$$

Hence, SPB assumption holds if

$$\tau > \frac{\sqrt{2}}{\mathrm{erf}(1)}\frac{\nu^3}{\mu\sigma^2}.$$

It remains to show that $\mathrm{erf}(1) > {}^1\!/\!\sqrt{2}$. Convexity of e^x on $x \in [-1, 0]$ implies $e^x \geq 1 + (1 - {}^1\!/\!e)x$ for any $x \in [-1, 0]$. Therefore

$$\begin{aligned}\mathrm{erf}(1) &= \frac{2}{\sqrt{\pi}} \int_0^1 e^{-t^2}\, dt \\ &\geq \frac{2}{\sqrt{\pi}} \int_0^1 \left(1 - (1 - {}^1\!/\!e)t^2\right) dt \\ &= \frac{2}{\sqrt{\pi}} \left(\frac{2}{3} + \frac{1}{3e}\right) > \frac{2}{\sqrt{4}} \left(\frac{2}{3} + \frac{1}{3 \cdot 3}\right) = \frac{7}{9} > \frac{1}{\sqrt{2}}.\end{aligned}$$

□

Lemma (3) follows from Lemma (12) applying it to i.i.d. data $\hat{g}_i^1(x), \hat{g}_i^2(x), \ldots, \hat{g}_i^M(x)$.

B.4 Sufficient conditions for SPB: Proof of Lemma 4

This observation is followed by the fact that for continuous random variables, the Gaussian distribution has the maximum differential entropy for a given variance[4]. Formally, let $p_G(x)$ be the probability density function (PDF) of a Gaussian random variable with variance σ^2 and $p(x)$ be the PDF of some random variable with the same variance. Then $H(p) \leq H(p_G)$, where

$$H(p) = -\int_{\mathbb{R}} p(x) \log p(x)\, dx$$

is the differential entropy of probability distribution $p(x)$ or alternatively differential entropy of random variable with PDF $p(x)$. Differential entropy for normal distribution can be expressed analytically by $H(p_G) = \frac{1}{2}\log(2\pi e\sigma^2)$. Therefore

$$H(p) \leq \frac{1}{2}\log(2\pi e\sigma^2)$$

for any distribution $p(x)$ with variance σ^2. Now, under the bounded variance assumption $\mathbb{E}\left[|\hat{g} - g|^2\right] \leq C$ (where g is the expected value of $\hat{g}$) we have the entropy of random variable $\hat{g}$ bounded by $\frac{1}{2}\log(2\pi eC)$. However, under the SPB assumption $\mathrm{Prob}\,(\mathrm{sign}\,\hat{g} = \mathrm{sign}\,g) > {}^1\!/\!2$ the entropy is unbounded. In order to prove this, it is enough to notice that under SPB assumption random variable $\hat{g}$ could be any Gaussian random variable with mean $g \neq 0$. In other words, SPB assumption holds for any Gaussian random variable with non-zero mean. Hence the entropy could be arbitrarily large as there is no restriction on the variance.

B.5 Convergence analysis for $M = 1$: Proof of Theorem 5

Basically, the analysis follows the standard steps used to analyze SGD for non-convex objectives, except the part (16)–(19) where inner product $\mathbb{E}[\langle g_k, \mathrm{sign}\,\hat{g}_k\rangle]$ needs to be estimated. This is exactly the place when stochastic gradient estimator $\mathrm{sign}\,\hat{g}_k$ interacts with the true gradient g_k. In case of standard SGD, we use estimator $\hat{g}_k$ and the mentioned inner product yields $\|g_k\|^2$, which is then used to measure the progress of the method. In our case, we show that

$$\mathbb{E}[\langle g_k, \mathrm{sign}\,\hat{g}_k\rangle] = \|g_k\|_\rho,$$

with the ρ-norm defined in Definition 1.

[4] see https://en.wikipedia.org/wiki/Differential_entropy or https://en.wikipedia.org/wiki/Normal_distribution#Maximum_entropy

Now we present the proof in more details. First, from L-smoothness assumption we have

$$\begin{aligned} f(x_{k+1}) &= f(x_k - \gamma_k \operatorname{sign} \hat{g}_k) \\ &\le f(x_k) - \langle g_k, \gamma_k \operatorname{sign} \hat{g}_k \rangle + \sum_{i=1}^d \frac{L_i}{2}(\gamma_k \operatorname{sign} \hat{g}_{k,i})^2 \\ &= f(x_k) - \gamma_k \langle g_k, \operatorname{sign} \hat{g}_k \rangle + \frac{d\bar{L}}{2}\gamma_k^2, \end{aligned}$$

where $g_k = g(x_k)$, $\hat{g}_k = \hat{g}(x_k)$, $\hat{g}_{k,i}$ is the i-th component of $\hat{g}_k$ and $\bar{L}$ is the average value of L_i's. Taking conditional expectation given current iteration x_k gives

$$\mathbb{E}[f(x_{k+1})|x_k] \le f(x_k) - \gamma_k \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] + \frac{d\bar{L}}{2}\gamma_k^2. \tag{15}$$

Using the definition of success probabilities ρ_i we get

$$\begin{aligned} \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] &= \langle g_k, \mathbb{E}[\operatorname{sign} \hat{g}_k] \rangle & (16) \\ &= \sum_{i=1}^d g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] = \sum_{\substack{1\le i\le d \\ g_{k,i}\ne 0}} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] & (17) \\ &= \sum_{\substack{1\le i\le d \\ g_{k,i}\ne 0}} g_{k,i} \left(\rho_i(x_k) \operatorname{sign} g_{k,i} + (1-\rho_i(x_k))(-\operatorname{sign} g_{k,i})\right) & (18) \\ &= \sum_{\substack{1\le i\le d \\ g_{k,i}\ne 0}} (2\rho_i(x_k) - 1)|g_{k,i}| = \sum_{i=1}^d (2\rho_i(x_k) - 1)|g_{k,i}| = \|g_k\|_\rho. & (19) \end{aligned}$$

Plugging this into (15) and taking full expectation, we get

$$\mathbb{E}\|g_k\|_\rho \le \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_k} + \frac{d\bar{L}}{2}\gamma_k. \tag{20}$$

Therefore

$$\sum_{k=0}^{K-1} \gamma_k \mathbb{E}\|g_k\|_\rho \le (f(x_0) - f^*) + \frac{d\bar{L}}{2}\sum_{k=0}^{K-1}\gamma_k^2. \tag{21}$$

Now, in case of decreasing step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$

$$\begin{aligned} \min_{0\le k<K} \mathbb{E}\|g_k\|_\rho &\le \sum_{k=0}^{K-1} \frac{\gamma_0}{\sqrt{k+1}} \mathbb{E}\|g_k\|_\rho \Big/ \sum_{k=0}^{K-1} \frac{\gamma_0}{\sqrt{k+1}} \\ &\le \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \frac{d\bar{L}}{2}\gamma_0 \sum_{k=0}^{K-1}\frac{1}{k+1}\right] \\ &\le \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \gamma_0 d\bar{L} + \frac{\gamma_0 d\bar{L}}{2}\log K\right] \\ &= \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \gamma_0 d\bar{L}\right] + \frac{\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}. \end{aligned}$$

where we have used the following standard inequalities

$$\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \geq \sqrt{K}, \quad \sum_{k=1}^{K} \frac{1}{k} \leq 2 + \log K. \tag{22}$$

In the case of constant step size $\gamma_k = \gamma$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\|_\rho \leq \frac{1}{\gamma K}\left[(f(x_0) - f^*) + \frac{d\bar{L}}{2}\gamma^2 K\right] = \frac{f(x_0) - f^*}{\gamma K} + \frac{d\bar{L}}{2}\gamma.$$

B.6 Convergence analysis for $M = 1$: Proof of Theorem 6

Clearly, the iterations $\{x_k\}_{k\geq 0}$ of Algorithm 1 with Option 2 do not increase the function value in any iteration, i.e. $\mathbb{E}[f(x_{k+1})|x_k] \leq f(x_k)$. Continuing the proof of Theorem 5 from (20), we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\|_\rho &\leq \frac{1}{K}\sum_{k=0}^{K-1} \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_k} + \frac{d\bar{L}}{2}\gamma_k \\
&= \frac{1}{K}\sum_{k=0}^{K-1} \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_0}\sqrt{k+1} + \frac{d\bar{L}}{2K}\sum_{k=0}^{K-1} \frac{\gamma_0}{\sqrt{k+1}} \\
&\leq \frac{1}{\sqrt{K}}\sum_{k=0}^{K-1} \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_0} + \frac{\gamma_0 d\bar{L}}{\sqrt{K}} \\
&= \frac{f(x_0) - \mathbb{E}[f(x_K)]}{\gamma_0\sqrt{K}} + \frac{\gamma_0 d\bar{L}}{\sqrt{K}} \\
&\leq \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \gamma_0 d\bar{L}\right],
\end{aligned}$$

where we have used the following inequality

$$\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \leq 2\sqrt{K}.$$

The proof for constant step size is the same as in Theorem 5.

B.7 Convergence analysis in parallel setting: Proof of Theorem 7

First, denote by $I(p; a, b)$ the regularized incomplete beta function, which is defined as follows

$$I(p; a, b) = \frac{B(p; a, b)}{B(a, b)} = \frac{\int_0^p t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}, \quad a, b > 0,\ p \in [0, 1]. \tag{23}$$

The proof of Theorem 7 goes with the same steps as in Theorem 5, except the derivation (16)–(19) is replaced by

$$\begin{aligned}
\mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k^{(M)} \rangle] &= \langle g_k, \mathbb{E}[\operatorname{sign} \hat{g}_k^{(M)}] \rangle \\
&= \sum_{i=1}^{d} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}^{(M)}] \\
&= \sum_{\substack{1 \le i \le d \\ g_{k,i} \neq 0}} |g_{k,i}| \cdot \mathbb{E}\left[\operatorname{sign}\left(\hat{g}_{k,i}^{(M)} \cdot g_{k,i}\right)\right] \\
&= \sum_{\substack{1 \le i \le d \\ g_{k,i} \neq 0}} |g_{k,i}| \left(2I(\rho_i(x_k); l, l) - 1\right) = \|g_k\|_{\rho_M},
\end{aligned}$$

where we have used the following lemma.

Lemma 13. *Assume that for some point $x \in \mathbb{R}^d$ and some coordinate $i \in \{1, 2, \dots, d\}$, master node receives M independent stochastic signs $\operatorname{sign} \hat{g}_i^n(x)$, $n = 1, \dots, M$ of true gradient $g_i(x) \neq 0$. Let $\hat{g}^{(M)}(x)$ be the sum of stochastic signs aggregated from nodes:*

$$\hat{g}^{(M)} = \sum_{n=1}^{M} \operatorname{sign} \hat{g}^n.$$

Then

$$\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(M)} \cdot g_i\right)\right] = 2I(\rho_i; l, l) - 1, \tag{24}$$

where $l = \lfloor \frac{M+1}{2} \rfloor$ and $\rho_i > 1/2$ is the success probablity for coordinate i.

Proof. Denote by S_i^n the Bernoulli trial of node n corresponding to ith coordinate, where "success" is the sign match between stochastic gradient and gradient:

$$S_i^n := \begin{cases} 1, & \text{if } \operatorname{sign} \hat{g}_i^n = \operatorname{sign} g_i \\ 0, & \text{otherwise} \end{cases} \sim \text{Bernoulli}(\rho_i). \tag{25}$$

Since nodes have their own independent stochastic gradients and the objective function (or dataset) is shared, then master node receives i.i.d. trials S_i^n, which sum up to a binomial random variable S_i:

$$S_i := \sum_{n=1}^{M} S_i^n \sim \text{Binomial}(M, \rho_i). \tag{26}$$

First, let us consider the case when there are odd number of nodes, i.e. $M = 2l - 1$, $l \geq 1$. In this case, taking into account (25) and (26), we have

$$\begin{aligned}
\text{Prob}\left(\operatorname{sign} \hat{g}_i^{(M)} = 0\right) &= 0, \\
\rho_i^{(M)} := \text{Prob}\left(\operatorname{sign} \hat{g}_i^{(M)} = \operatorname{sign} g_i\right) &= \text{Prob}(S_i \geq l), \\
1 - \rho_i^{(M)} = \text{Prob}\left(\operatorname{sign} \hat{g}_i^{(M)} = -\operatorname{sign} g_i\right)&.
\end{aligned}$$

It is well known that cumulative distribution function of binomial random variable can be expressed with regularized incomplete beta function:

$$\text{Prob}(S_i \geq l) = I(\rho_i; l, M - l + 1) = I(\rho_i; l, l). \tag{27}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(M)}\cdot g_i\right)\right] &= \rho_i^{(M)}\cdot 1 + (1-\rho_i^{(M)})\cdot(-1)\\ &= 2\rho_i^{(M)}-1\\ &= 2\mathrm{Prob}(S_i\ge l)-1\\ &= 2I(\rho_i;l,l)-1.\end{aligned}$$

In the case of even number of nodes, i.e. $M=2l$, $l\ge 1$, there is a probability to fail the vote $\mathrm{Prob}\left(\operatorname{sign}\hat{g}_i^{(M)}=0\right)>0$. However using (27) and properties of beta function[5] gives

$$\begin{aligned}\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(2l)}\cdot g_i\right)\right] &= \mathrm{Prob}(S_i\ge l+1)\cdot 1+\mathrm{Prob}(S_i\le l-1)\cdot(-1)\\ &= I(\rho_i;l+1,l)+I(\rho_i;l,l+1)-1\\ &= 2I(\rho_i;l,l)-1\\ &= \mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(2l-1)}\cdot g_i\right)\right].\end{aligned}$$

This also shows that in expectation there is no difference between having $2l-1$ and $2l$ nodes. □

B.8 Convergence analysis in parallel setting: Speedup with respect to M

Here we present the proof of exponential noise reduction in parallel setting in terms of number of nodes. We first state the well-known Hoeffding's inequality:

Theorem 14 (Hoeffding's inequality for general bounded random variables; see [Vershynin, 2018], Theorem 2.2.6)**.** *Let $X_1, X_2, \ldots, X_M$ be independent random variables. Assume that $X_n\in[A_n,B_n]$ for every n. Then, for any $t>0$, we have*

$$\mathrm{Prob}\left(\sum_{n=1}^M (X_n-\mathbb{E}X_n)\ge t\right)\le \exp\left(-\frac{2t^2}{\sum_{n=1}^M (B_n-A_n)^2}\right).$$

Define random variables X_i^n, $n=1,2,\ldots,M$ showing the missmatch between stochastic gradient sign and full gradient sign from node n and coordinate i:

$$X_i^n := \begin{cases}-1, & \text{if } \operatorname{sign}\hat{g}_i^n=\operatorname{sign} g_i\\ 1, & \text{otherwise}\end{cases} \tag{28}$$

Clearly $\mathbb{E}X_i^n = 1-2\rho_i$ and Hoeffding's inequality gives

$$\mathrm{Prob}\left(\sum_{n=1}^M X_i^n - M(1-2\rho_i)\ge t\right)\le\exp\left(-\frac{t^2}{2M}\right),\quad t>0.$$

Choosing $t=M(2\rho_i-1)>0$ (because of SPB assumption) yields

$$\mathrm{Prob}\left(\sum_{n=1}^M X_i^n\ge 0\right)\le\exp\left(-\frac{1}{2}(2\rho_i-1)^2M\right).$$

[5] see `https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function`

Using Lemma 24, we get

$$2I(\rho_i, l; l) - 1 = \mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(M)} \cdot g_i\right)\right] = 1 - \operatorname{Prob}\left(\sum_{n=1}^{M} X_i^n \geq 0\right) \geq 1 - \exp\left(-(2\rho_i - 1)^2 l\right),$$

which provides the following estimate for ρ_M-norm:

$$\left(1 - \exp\left(-(2\rho(x) - 1)^2 l\right)\right) \|g(x)\|_1 \leq \|g(x)\|_{\rho_M} \leq \|g(x)\|_1,$$

where $\rho(x) = \min_{1 \leq i \leq d} \rho_i(x) > \frac{1}{2}$.

B.9 Distributed training with partitioned data: Proof of Theorem 8

We follow the analysis of Cutkosky and Mehta [2020], who derived similar convergence rate for normalized SGD in single node setting. The novelty in our proof technique is i) extending the analysis in distributed setting and ii) establishing a connection between normalized SGD and sign-based methods via the new notion of *stochastic sign.*

Lemma 15 (see Lemma 2 in [Cutkosky and Mehta, 2020])**.** *For any non-zero vectors a and b*

$$-\frac{\langle a, b\rangle}{\|a\|} \leq -\frac{1}{3}\|b\| + \frac{8}{3}\|a - b\|.$$

Proof. Denote $c = a - b$ and consider two cases. If $\|c\| \leq \frac{1}{2}\|b\|$, then

$$-\frac{\langle a, b\rangle}{\|a\|} = -\frac{\|b\|^2 + \langle c, b\rangle}{\|a\|} \leq -\frac{\|b\|^2 - \|c\|\|b\|}{\|b + c\|} \leq -\frac{\|b\|^2}{2\|b + c\|} \leq -\frac{1}{3}\|b\| \leq -\frac{1}{3}\|b\| + \frac{8}{3}\|a - b\|.$$

Alternatively, if $\|c\| > \frac{1}{2}\|b\|$, then

$$-\frac{\langle a, b\rangle}{\|a\|} \leq \|b\| \leq -\frac{1}{3}\|b\| + \frac{4}{3}\|b\| \leq -\frac{1}{3}\|b\| + \frac{8}{3}\|c\|.$$

□

We start from the smoothness of functions f_n

$$\begin{aligned}
f(x_{k+1}) &= \frac{1}{M}\sum_{n=1}^{M} f_n\left(x_k - \frac{\gamma}{M}s_k\right) \\
&\leq \frac{1}{M}\sum_{n=1}^{M}\left[f_n(x_k) - \frac{\gamma}{M}\langle \nabla f_n(x_k), s_k\rangle + \frac{L_n\gamma^2}{2M^2}\|s_k\|^2\right] \\
&= f(x_k) - \frac{\gamma}{M}\langle \nabla f(x_k), s_k\rangle + \frac{\tilde{L}\gamma^2}{2}\left\|\frac{1}{M}\sum_{n=1}^{M} s_k^n\right\|^2 \\
&\leq f(x_k) - \frac{\gamma}{M}\langle \nabla f(x_k), s_k\rangle + \frac{\tilde{L}\gamma^2}{2}\frac{1}{M}\sum_{n=1}^{M}\|s_k^n\|^2 \\
&= f(x_k) - \frac{\gamma}{M}\sum_{n=1}^{M}\langle \nabla f(x_k), s_k^n\rangle + \frac{d\tilde{L}\gamma^2}{2}.
\end{aligned}$$

Denote $g_k^n = \nabla f_n(x_k)$, $g_k = \nabla f(x_k)$. Taking expectation conditioned on previous iterate x_k and current stochastic gradient $\hat{g}_k^n$, we get

$$\begin{aligned}\mathbb{E}\left[f(x_{k+1})|x_k, \hat{g}_k^n\right] &\le f(x_k) - \frac{\gamma}{M}\sum_{n=1}^{M}\frac{\langle g_k, m_k^n\rangle}{\|m_k^n\|} + \frac{d\tilde{L}\gamma^2}{2} \\ &\overset{\text{Lemma 15}}{\le} f(x_k) - \frac{\gamma}{3}\|g_k\| + \frac{8\gamma}{3M}\sum_{n=1}^{M}\|m_k^n - g_k\| + \frac{d\tilde{L}\gamma^2}{2}.\end{aligned} \tag{29}$$

Next, we find recurrence relation for the error terms $\hat{\epsilon}_k^n \coloneqq m_k^n - g_k$. Denote $\epsilon_k^n \coloneqq \hat{g}_k^n - g_k$, and notice that

$$\begin{aligned}\hat{\epsilon}_{k+1}^n &= \beta m_k^n + (1-\beta)\hat{g}_{k+1}^n - g_{k+1} \\ &= \beta(m_k^n - g_{k+1}) + (1-\beta)(\hat{g}_{k+1}^n - g_{k+1}) \\ &= \beta(m_k^n - g_k) + \beta(g_k - g_{k+1}) + (1-\beta)\epsilon_{k+1}^n \\ &= \beta\hat{\epsilon}_k^n + \beta(g_k - g_{k+1}) + (1-\beta)\epsilon_{k+1}^n.\end{aligned}$$

Unrolling this recursion and noting that $\hat{\epsilon}_0^n = \epsilon_0^n$ (due to initial moment $m_{-1}^n = \hat{g}_0^n$), we get

$$\hat{\epsilon}_{k+1}^n = \beta^{k+1}\epsilon_0^n + \beta\sum_{t=0}^{k}\beta^t(g_{k-t} - g_{k+1-t}) + (1-\beta)\sum_{t=0}^{k}\beta^t\epsilon_{k+1-t}^n$$

From Assumption 2, we have

$$\mathbb{E}\left[\langle\epsilon_k^n, \epsilon_{k'}^n\rangle\right]\begin{cases}\le(\sigma^n)^2 & \text{if } k = k', \\ = 0 & \text{if } k \ne k'.\end{cases} \tag{30}$$

Using L^n-smoothness of functions f_n again, we have

$$\|g_k - g_{k+1}\| \le \frac{1}{M}\sum_{n=1}^{M}\|g_k^n - g_{k+1}^n\| \le \frac{1}{M}\sum_{n=1}^{M}L^n\|x_k - x_{k+1}\| = \frac{\tilde{L}\gamma}{M}\|s_k\| \le \tilde{L}\gamma\sqrt{d}. \tag{31}$$

Therefore

$$\begin{aligned}\mathbb{E}\|\hat{\epsilon}_{k+1}^n\| &\le \beta^{k+1}\|\epsilon_0^n\| + \sum_{t=0}^{k}\beta^{t+1}\|g_{k-t} - g_{k+1-t}\| + (1-\beta)\mathbb{E}\left\|\sum_{t=0}^{k}\beta^t\epsilon_{k+1-t}^n\right\| \\ &\overset{(31)}{\le} \beta^{k+1}\sigma^n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + (1-\beta)\sqrt{\mathbb{E}\left\|\sum_{t=0}^{k}\beta^t\epsilon_{k+1-t}^n\right\|^2} \\ &\overset{(30)}{\le} \beta^{k+1}\sigma^n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + (1-\beta)\sqrt{\sum_{t=0}^{k}\beta^{2t}(\sigma^n)^2} \\ &\le \beta^{k+1}\sigma^n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + \sigma^n\sqrt{1-\beta}\end{aligned}$$

Averaging this bound over the nodes yields

$$\frac{1}{M}\sum_{n=1}^{M}\mathbb{E}\|\hat{\epsilon}_k^n\| \le \beta^k\tilde{\sigma} + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + \tilde{\sigma}\sqrt{1-\beta}.$$

Then averaging over the iterates gives

$$\frac{1}{KM}\sum_{k=0}^{K-1}\sum_{n=1}^{M}\mathbb{E}\|\hat{\epsilon}_k^n\| \le \frac{\tilde{\sigma}}{(1-\beta)K} + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + \tilde{\sigma}\sqrt{1-\beta}.$$

Taking full expectation in (29), we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|g_k\| &\le \frac{3}{\gamma K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x_k)-f(x_{k+1})\right] + \frac{8}{MK}\sum_{n=1}^{M}\sum_{k=0}^{K-1}\mathbb{E}\|\hat{\epsilon}_k^n\| + \frac{3}{2}\tilde{L}d\gamma \\ &\le \frac{3(f(x_0)-f_*)}{\gamma K} + \frac{8\tilde{\sigma}}{(1-\beta)K} + \frac{8\tilde{L}\gamma\sqrt{d}}{1-\beta} + 8\tilde{\sigma}\sqrt{1-\beta} + \frac{3}{2}\tilde{L}d\gamma.\end{aligned}$$

Now it remains to choose parameters γ and β properly. Setting $\beta = 1-\frac{1}{\sqrt{K}}$ and $\gamma = \frac{1}{K^{3/4}}$, we get

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|g_k\| &\le \frac{3\delta_f}{\gamma K} + \frac{8\tilde{\sigma}}{(1-\beta)K} + \frac{8\tilde{L}\gamma\sqrt{d}}{1-\beta} + 8\tilde{\sigma}\sqrt{1-\beta} + 3\tilde{L}d\gamma \\ &\le \frac{1}{K^{1/4}}\left[3\delta_f + 16\tilde{\sigma} + 8\tilde{L}\sqrt{d} + \frac{3\tilde{L}d}{\sqrt{K}}\right].\end{aligned}$$

C Recovering Theorem 1 in [Bernstein et al., 2019] from Theorem 5

To recover Theorem 1 in [Bernstein et al., 2019], first note that choosing a particular step size γ in (8) yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|g_k\|_\rho \le \sqrt{\frac{2d\bar{L}(f(x_0)-f^*)}{K}}, \quad \text{with} \quad \gamma = \sqrt{\frac{2(f(x_0)-f^*)}{d\bar{L}K}}. \tag{32}$$

Then, due to Lemma 1, under unbiasedness and unimodal symmetric noise assumption, we can lower bound general ρ-norm by mixed $l^{1,2}$ norm. Finally we further lower bound our $l^{1,2}$ norm to obtain *the mixed norm* used in Theorem 1 of [Bernstein et al., 2019]:

$$\begin{aligned}5\sqrt{\frac{d\bar{L}(f(x_0)-f^*)}{K}} &\ge \frac{5}{\sqrt{2}}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|g_k\|_\rho \\ &\ge \frac{5}{\sqrt{2}}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|g_k\|_{l^{1,2}} = \frac{5}{\sqrt{2}}\frac{1}{K}\sum_{k=0}^{K-1}\left[\sum_{i=1}^{d}\frac{g_i^2}{|g_i|+\sqrt{3}\sigma_i}\right] \\ &\ge \frac{5}{\sqrt{2}}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{2}{5}\sum_{i\in H_k}|g_{k,i}| + \frac{\sqrt{3}}{5}\sum_{i\notin H_k}\frac{g_{k,i}^2}{\sigma_i}\right] \\ &\ge \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sum_{i\in H_k}|g_{k,i}| + \sum_{i\notin H_k}\frac{g_{k,i}^2}{\sigma_i}\right],\end{aligned}$$

where $H_k = \{1 \le i \le d \colon \sigma_i < \frac{\sqrt{3}}{2}|g_{k,i}|\}$.