On Stochastic Sign Descent Methods

Mher Safaryan MHER.SAFARYAN.1@KAUST.EDU.SA
KAUST, KSA

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
KAUST, KSA & MIPT, Russia

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 [2], have recently been gaining popularity because of their simple compression rule and connection to adaptive gradient methods, like ADAM. In this paper, we provide a unified and general analysis of sign-based methods for non-convex optimization. Our analysis (i) is built on intuitive bounds on success probabilities allowing biased estimators, (ii) does not rely on special noise distributions nor on the boundedness of the variance of stochastic gradients, (iii) recovers existing convergence rates as a special case and (iv) extending the theory to distributed setting within a parameter server framework, we guarantee exponentially fast variance reduction with respect to number of nodes, maintaining 1-bit compression in both directions and using small mini-batch sizes. (v) We also discuss a simple idea to fix the convergence issues of signSGD. Finally, we validate our theoretical findings experimentally.

1. Introduction

One of the key factors behind the success of modern machine learning models is the availability of large amounts of training data [4, 9, 12]. 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

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

where $f_m : \mathbb{R}^d \to \mathbb{R}$ represents the *non-convex* loss of a deep learning model parameterized by $x \in \mathbb{R}^d$ associated with data stored on node m. Arguably, stochastic gradient descent (SGD) [10, 11, 15] in of its many variants [5, 6, 8, 13, 17] is the most popular algorithm for solving (1). In its basic implementation, all workers $m \in \{1, 2, \ldots, M\}$ in parallel compute a random approximation $\hat{g}^m(x_k)$ of $\nabla f_m(x_k)$, known as the *stochastic gradient*. These approximations are then sent to a master node which performs the aggregation $\hat{g}(x_k) := \frac{1}{M}\sum_{m=1}^{M} \hat{g}^m(x_k)$. 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 the theoretical results obtained in this work. $\tilde{\mathcal{O}}$ notation ignores logarithmic factors and $\mathcal{O}^*$ notation shows the rate to a neighbourhood of the solution.

	This work Theorem 2	**This work** Theorem 13	SGD (Sec. D) Theorem 16	signSGD, [3] Theorem 1
Step size $\gamma_k \equiv \gamma$	$\mathcal{O}^*\left(\frac{1}{K}\right)$	$\mathcal{O}^*\left(\frac{1}{K}\right)$	$\mathcal{O}^*\left(\frac{1}{K}\right)$	✗
Step size $\gamma_k = \frac{\gamma_0}{\sqrt{k+1}}$	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right)$	✗
Step size $\gamma = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$
Can handle biased estimators?	✓	✓	✗	✗
Weak dependence on smoothness parameters?	✓ $\frac{1}{d}\sum_{i=1}^{d} L_i$	✓ $\frac{1}{d}\sum_{i=1}^{d} L_i$	✗ $\max_{i=1}^{d} L_i$	✓ $\frac{1}{d}\sum_{i=1}^{d} L_i$
Weak noise assumptions?	✓ $\rho_i > \frac{1}{2}$	✓ $\rho_i > \frac{1}{2}$	✗ $\mathbb{E}\\|\hat{g}\\|_2^2 \leq C$ [1]	✗ unimodal, symmetric & $\mathrm{Var}[\hat{g}_i] \leq \sigma_i^2$
Gradient norm used in theory	ρ-norm	ρ-norm	(squared) l^2	A mix of l^1 and l^2 norms

1.1. Contributions

We now summarize the main contributions of this work.

• **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[2]

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

while the second has a new form not considered in the literature before (see Section C.6):

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

• **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 $\nabla f(x^k)$ with a probability strictly larger than $1/2$ (see Assumption 1). We 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 2 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.

2. $\operatorname{sign} g$ is applied element-wise to the entries $g_1, g_2, \dots, g_d$ 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 $t < 0$.

Finally, Theorem 1 in [3] can be recovered from our general Theorem 2. 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.

• **Distributed setup.** We extend our results to the *distributed setting* with arbitrary M, where we also consider sign-based compression of the aggregated gradients. We guarantee exponentially fast variance reduction with respect to the number of nodes.

• **A simple fix.** We introduce *stochastic sign*, which fixes the convergence issues of SIGNSGD.

2. Success Probabilities and Gradient Noise

In this section we describe our key (and weak) assumption on the gradient estimator $\hat{g}(x)$, and give an example which shows that without this assumption, SIGNSGD can fail to converge.

Assumption 1 (SPB: Success Probability Bounds) *For any $x \in \mathbb{R}^d$, we have 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 for all $x \in \mathbb{R}^d$ and all $i \in \{1, 2, \dots, d\}$*

$$\rho_i(x) := \mathrm{Prob}\left(\mathrm{sign}\, \hat{g}_i(x) = \mathrm{sign}\, g_i(x)\right) > \tfrac{1}{2}. \tag{4}$$

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. We stress that *Assumption 1* is the *only* assumption on gradient noise in this paper. 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 $\rho_i \approx \frac{1}{2}$.

• **2.1. A counterexample to SIGNSGD.** Here we analyze a counterexample to signSGD discussed in [7]. Consider the following least-squares problem with unique minimizer $x^* = (0, 0)$:

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

where $\varepsilon \in (0, 1)$ and $\hat{g}(x) = \nabla\langle a_i, x\rangle^2 = 2\langle a_i, x\rangle a_i$ with probabilities $1/2$. Then $\mathrm{sign}\, \hat{g}(x) = (-1)^i \,\mathrm{sign}\langle a_i, x\rangle \left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ with probabilities $1/2$. Notice that signSGD with any step-size remains stuck since $x \in H := \{(z_1, z_2) \colon z_1 + z_2 = 2\}$ implies $\mathrm{sign}\, \hat{g}(x) = \pm(1, -1) \parallel H$. In this case, SPB assumption (4) is violated as $\rho_i(x) \le 1/2$ when $\langle a_1, x\rangle \cdot \langle a_2, x\rangle > 0$.

• **2.2. 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 $g(x) = \nabla f(x)$.

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

2. Let coordinate-wise variances $\sigma_i^2(x) \le c_i\, g_i^2(x)$ are bounded for some constants c_i. Choose mini-batch size $\tau > 2\max_i c_i$. If further $g_i \neq 0$, then $\rho_i \ge 1 - \frac{c_i}{\tau} > \frac{1}{2}$.

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

4. Differential entropy of a probability distribution under the bounded variance assumption is bounded, while under the SPB assumption it could be arbitrarily large.

We introduce a norm-like function ρ-*norm*, induced from success probabilities and used to measure gradients in our convergence rates (see Section A).

Definition 1 (ρ-norm) *Let $\rho := \{\rho_i(x)\}_{i=1}^d$ be the collection of probability functions from the SPB assumption. We define the ρ-norm of gradient $g(x)$ via $\|g(x)\|_\rho := \sum_{i=1}^d (2\rho_i(x)-1)|g_i(x)|$.*

3. Convergence Theory

Throughout the paper we assume that nonconvex $f\colon \mathbb{R}^d \to \mathbb{R}$ is lower bounded, i.e., $f(x) \geq f^*$ for all $x \in \mathbb{R}^d$, and L-smooth with some non-negative constants $L = (L_i)_{i=1}^d$, i.e., $f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \sum_{i=1}^d \frac{L_i}{2}(y_i-x_i)^2$ for all $x,\ y \in \mathbb{R}^d$. Let $\bar{L} := \frac{1}{d}\sum_i L_i$ and $L_{\max} := \max_i L_i$.

Algorithm 1 SIGNSGD

1: **Input:** step size γ_k, current point x_k
2: $\hat{g}_k \leftarrow \text{StochasticGradient}(x_k)$
3: $x_{k+1} \leftarrow x_k - \gamma_k \operatorname{sign} \hat{g}_k$

• **3.1. Convergence Analysis for $M=1$.** We now state our convergence result for signSGD (2).

Theorem 2 (Non-convex convergence of signSGD) *Under the SPB assumption, signSGD (Algorithm 1) with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as follows*

$$\min_{0\leq k<K} \mathbb{E}\|\nabla f(x_k)\|_\rho \leq \frac{1}{\sqrt{K}}\left[\frac{f(x_0)-f^*}{\gamma_0} + \gamma_0 d\bar{L}\right] + \frac{\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}\,. \tag{5}$$

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

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_\rho \leq \frac{f(x_0)-f^*}{\gamma K} + \frac{\gamma d\bar{L}}{2}\,. \tag{6}$$

• **Generalization.** Theorem 2 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 [2, 3] 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 [3] can be recovered from Theorem 2 (see Section E).

• **Convergence rate.** Rates (5) and (6) 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 from (5) we get the rate $\tilde{\mathcal{O}}(1/\sqrt{K})$ 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.

• **3.2. Convergence Analysis in Distributed Setting.** In this part we present the convergence result of distributed signSGD (*Algorithm 2*) with majority vote introduced in [2]. 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.

Algorithm 2 DISTRIBUTED SIGNSGD WITH MAJORITY VOTE

1: **Input:** step sizes $\{\gamma_k\}$, current point x_k, # of nodes M
2: **on** each node
3: $\quad \hat{g}^m(x_k) \leftarrow \text{StochasticGradient}(x_k)$
4: **on** server
5: $\quad$ **pull** sign $\hat{g}^m(x_k)$ **from** each node
6: $\quad$ **push** sign $\left[\sum_{m=1}^M \operatorname{sign} \hat{g}^m(x_k)\right]$ **to** each node
7: **on** each node
8: $\quad x_{k+1} \leftarrow x_k - \gamma_k \operatorname{sign}\left[\sum_{m=1}^M \operatorname{sign} \hat{g}^m(x_k)\right]$

Before presenting the convergence rate, we briefly discuss the case when dataset is partitioned among the nodes in distributed setup. In this case each machine m has its own loss function $f_m(x)$ for $m = 1, 2, \ldots, M$. Under this setting even signGD (with full-batch gradients and no noise) fails to converge. Indeed, if we multiply each loss function $f_m(x)$ of m-th node by an arbitrary positive scalars $w_m > 0$, then the overall loss function $f_w(x) = \frac{1}{M}\sum_{i=m}^M w_m f_m(x)$ changes arbitrarily while the iterates of signGD are not changed at all as the master aggregates the same signs $\operatorname{sign}(w_m \nabla f_m(x)) = \operatorname{sign} \nabla f_m(x)$ no matter what w_m factors are. Thus, the convergence of distributed signSGD is considered in the case when all machines have full access to dataset.

Known convergence results [2, 3] use $O(K)$ mini-batch size as well as $O(1/K)$ constant step size. In the sequel we remove this limitations extending Theorem 2 to distributed training. In distributed setting 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 for the details).

Definition 3 (ρ_M-norm) *Let $M \ge 1$ be the number of nodes and $l = \left[\frac{M+1}{2}\right]$. Define ρ_M-norm of gradient $g(x)$ at $x \in \mathbb{R}^d$ via $\|g(x)\|_{\rho_M} := \sum_{i=1}^d \left(2I(\rho_i(x); l, l) - 1\right)|g_i(x)|$*

Now we can state the convergence rate of distributed signSGD with majority vote.

Theorem 4 (Non-convex convergence of distributed signSGD) *Under SPB assumption, distributed signSGD (Algorithm 2) with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as*

$$\min_{0\le k<K} \mathbb{E}\|\nabla f(x_k)\|_{\rho_M} \le \frac{1}{\sqrt{K}}\left[\frac{f(x_0)-f^*}{\gamma_0} + \gamma_0 d\bar{L}\right] + \frac{\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}. \tag{7}$$

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

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

Using Hoeffding's inequality, we show 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 $\rho(x) = \min_{1\le i\le d}\rho_i(x) > 1/2$. Hence, we have exponential variance reduction in terms of number of nodes.

• **3.3. A Simple Fix to signSGD: Stochastic signSGD.** The issue with signSGD is that sign compression of stochastic gradient is biased estimator, which also complicates the analysis. One way to overcome SPB assumption and make signSGD to converge in general is to incorporate a scaling factor together with error feedback mechanism [7], which can handle biased compressions. Another, and more natural, way of fixing the convergence issue is to introduce stochastic $\widetilde{\mathrm{sign}}$ operator.

Definition 5 (Stochastic Sign) *Let $\|\cdot\|$ be any norm. Define the stochastic sign operator $\widetilde{\mathrm{sign}} : \mathbb{R}^d \to \mathbb{R}^d$ via*

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

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

Stochastic $\widetilde{\mathrm{sign}}$ operator unlike the deterministic sign operator is unbiased with the scaling factor $\|g\|$. Indeed, if $\hat{g}$ is an unbiased estimator of g, i.e. $\mathbb{E}[\hat{g}] = g$, then

$$\mathbb{E}\left[\|\hat{g}\|\,\widetilde{\mathrm{sign}}\,\hat{g}\right] = \mathbb{E}\left[\mathbb{E}\left[\|\hat{g}\|\,\widetilde{\mathrm{sign}}\,\hat{g} \,|\, \hat{g}\right]\right] = \mathbb{E}\left[\|\hat{g}\|\left(\frac{1}{2} + \frac{1}{2}\frac{\hat{g}}{\|\hat{g}\|}\right) - \|\hat{g}\|\left(\frac{1}{2} - \frac{1}{2}\frac{\hat{g}}{\|\hat{g}\|}\right)\right] = g.$$

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

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

Appendix: "On Stochastic Sign Descent Methods"

Appendix A. 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 6 (ρ-norm) *Let $\rho := \{\rho_i(x)\}_{i=1}^d$ be the collection of probability functions from the SPB assumption. We define the ρ-norm of gradient $g(x)$ via*

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

Note that 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 $g = 0$. Under the assumptions of Lemma 2, ρ-norm can be lower bounded by a weighted l^1 norm with positive constant weights $1 - {}^{2c_i}/_{\tau} > 0$: $\|g\|_\rho = \sum_{i=1}^d (2\rho_i - 1)|g_i| \ge \sum_{i=1}^d (1 - {}^{2c_i}/_{\tau})|g_i|$. Under the assumptions of Lemma 1, ρ-norm can be lower bounded by a mixture of the l^1 and squared l^2 norms:

$$\|g\|_\rho = \sum_{i=1}^d (2\rho_i - 1)|g_i| \ge \sum_{i=1}^d \tfrac{g_i^2}{|g_i| + \sqrt{3}\sigma_i} := \|g\|_{l^{1,2}}. \tag{9}$$

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: i) $\|g\|_{l^{1,2}} \ge 0$ and $\|g\|_{l^{1,2}} = 0$ if and only if $g = 0$; ii) $g^k \to g$ (in l^2 sense) implies $\|g^k\|_{l^{1,2}} \to \|g\|_{l^{1,2}}$, and iii) $0 \le g_i \le \tilde{g}_i$ 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: $\frac{1}{(1+\sqrt{3}c)\varepsilon + \sqrt{3}\tilde{c}}\|g\|_2^2 \le \|g\|_{l^{1,2}} \le \frac{1}{\sqrt{3}\tilde{c}}\|g\|_2^2$. 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 in [3].

Appendix B. Experiments

We verify 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:

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

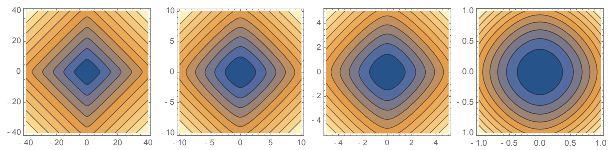

Figure 1: Contour plots of the $l^{1,2}$ norm (9) at 4 different scales with fixed noise $\sigma = 1$.

B.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 4 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 2 shows the robustness of SPB assumption in the convergence rate (6) 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.

B.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 3 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 3. 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 [1], 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 3), 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 2: 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.

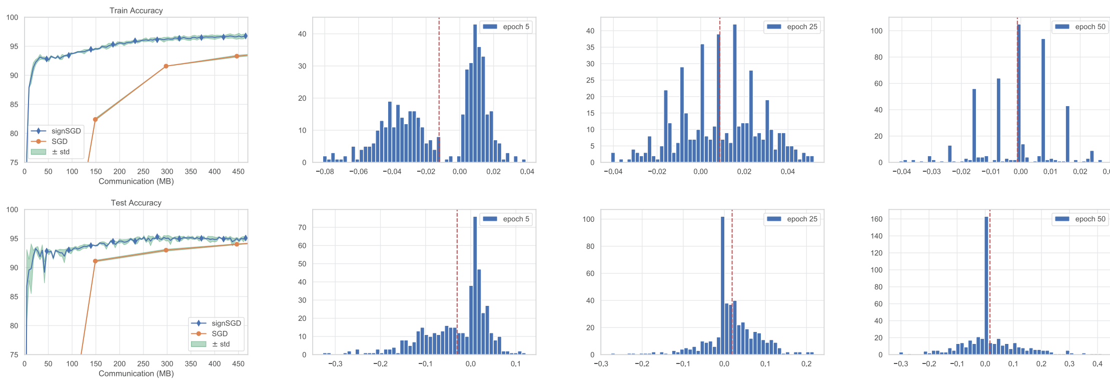

Figure 3: Convergence of signSGD and comparison with SGD on the MNIST dataset using the split batch construction strategy. The budget of gradient communication (MB) is fixed and the network is a single hidden layer FNN. We first tuned the constant step size over logarithmic scale $\{1, 0.1, 0.01, 0.001, 0.0001\}$ 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.

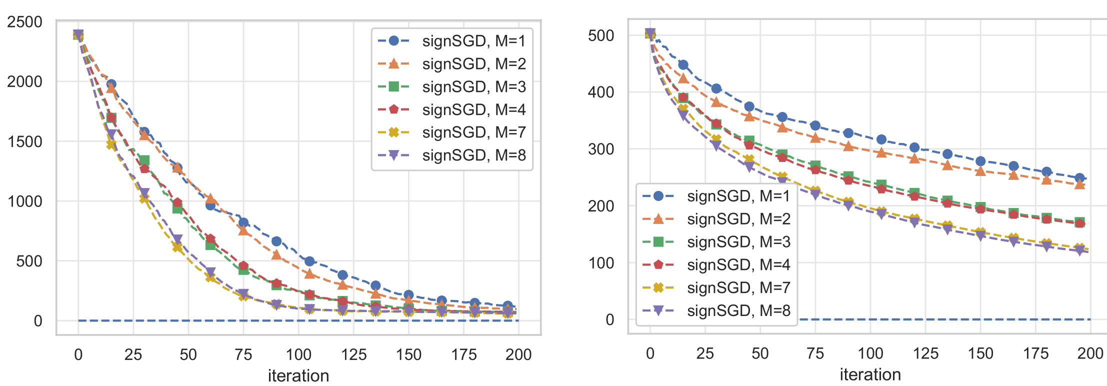

Figure 4: Experiments on distributed signSGD with majority vote using Rosenbrock function. Plots show function values with respect to iterations averaged over 10 repetitions. Left plot used constant step size $\gamma = 0.02$, right plot used variable step size with $\gamma_0 = 0.02$. We set mini-batch size 1 and used the same initial point. Dashed blue lines mark the minimum.

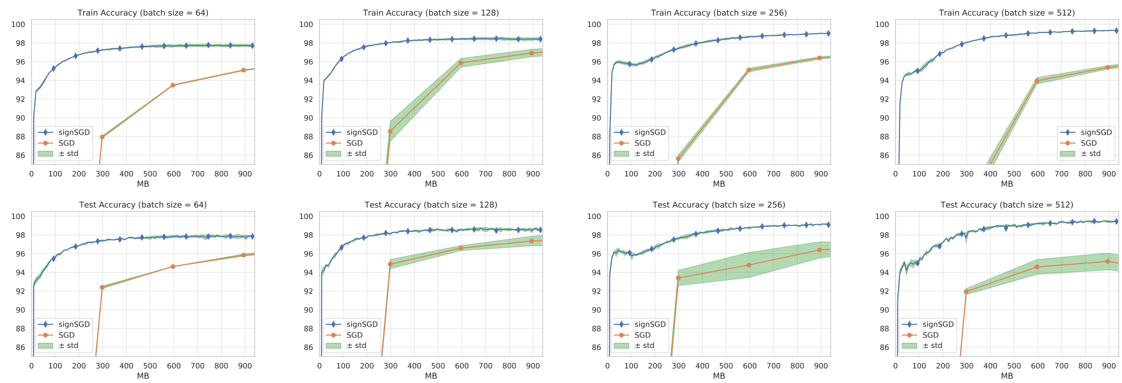

Figure 5: Comparison of signSGD and SGD on the MNIST dataset with a fixed budget of gradient communication (MB) using single hidden layer FNN and the standard batch construction strategy. For each batch size, we first tune the constant step size over logarithmic scale $\{10, 1, 0.1, 0.01, 0.001\}$ and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

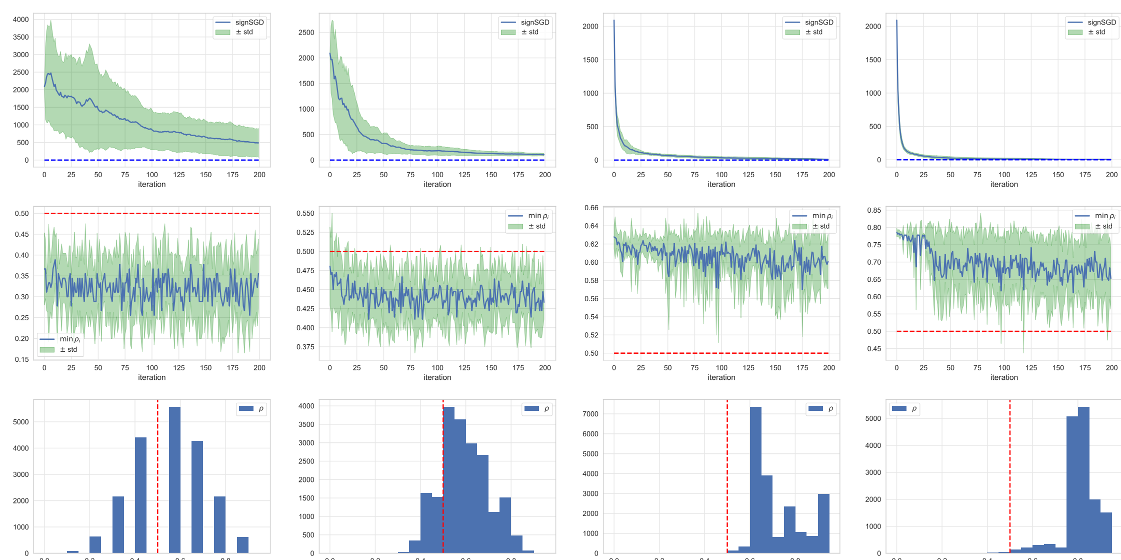

Figure 6: Performance of signSGD with variable step size ($\gamma_0 = 0.25$) under four different noise levels (mini-batch size 1, 2, 5, 7) using Rosenbrock function. As in the experiments of Figure 2 with constant step size, these plots show the relationship between success probabilities and the convergence rate (5). 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.

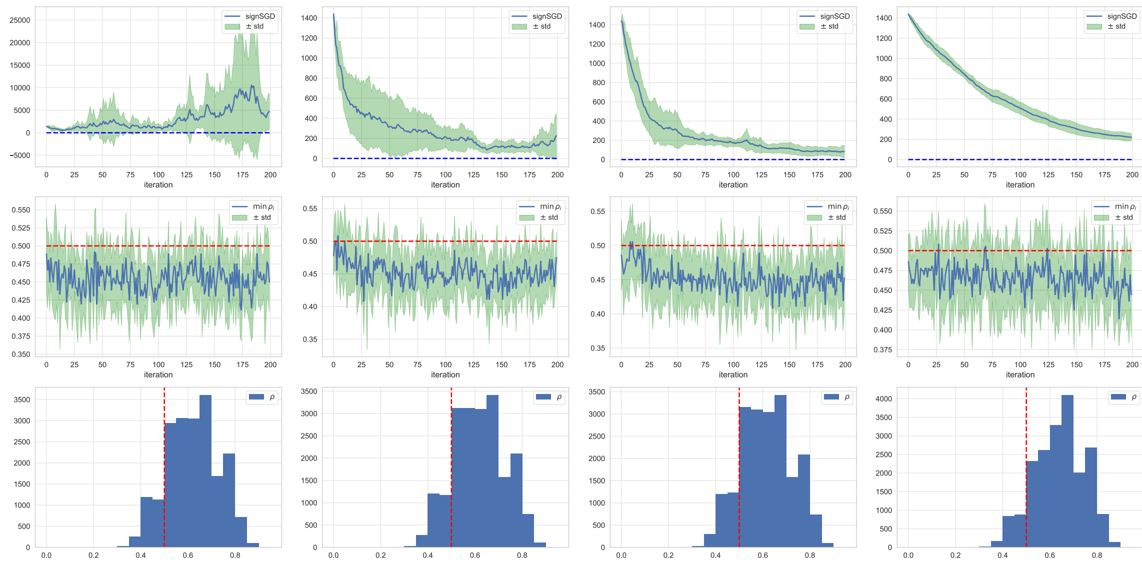

Figure 7: In this part of experiments we investigated convergence rate (6) to a neighborhood of the solution. We fixed gradient noise level by setting mini-batch size 2 and altered the constant step size. For the first column we set bigger step size $\gamma = 0.25$ to detect the divergence (as we slightly violated SPB assumption). Then for the second and third columns we set $\gamma = 0.1$ and $\gamma = 0.05$ to expose the convergence to a neighborhood of the minimizer. For the forth column we set even smaller step size $\gamma = 0.01$ to observe a slower convergence.

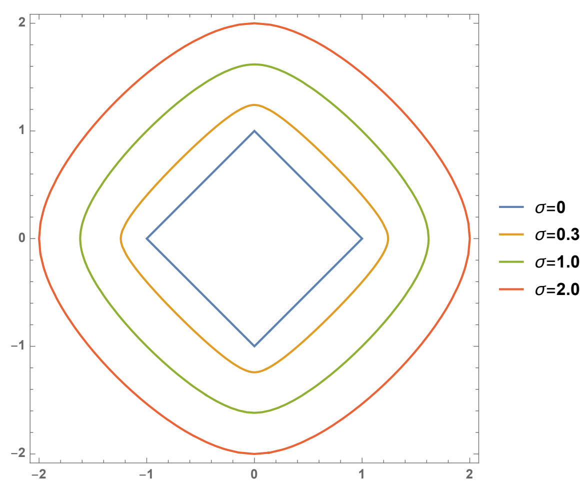

Figure 8: Unit balls in $l^{1,2}$ norm (9) with different noise levels.

Appendix C. Proofs

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

Here we state the well-known Gauss's inequality on unimodal distributions[3].

Theorem 7 (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,*

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

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

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

where $m=\mu$ and $\sigma_m^2=\sigma^2$ are the mean and variance of unimodal, symmetric random variable X, and $r\geq 0$. 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

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

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

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

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

3. see https://en.wikipedia.org/wiki/Gauss%27s_inequality

tighter as

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

Hence, continuing the proof of Lemma 1, we get

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

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

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

C.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.

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

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

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

which imples

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

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

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

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau}X_i\right] = \mathrm{sign}\,\mu\right) \ge 1-\frac{\sigma^2}{\tau\mu^2}. \tag{10}$$

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 $X_1, X_2, \ldots, X_\tau$. ■

Obviously, bound (10) 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 finitness of 3th central moment.

Lemma 9 *Let $X_1, X_2, \ldots, X_\tau$ be i.i.d. random variables with non-zero mean $\mu := \mathbb{E}X_1 \neq 0$, 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 $\tau \geq 1$*

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu\right) \geq \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{|\mu|\sqrt{\tau}}{\sqrt{2}\sigma}\right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right), \tag{11}$$

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[4] on the rate of approximation in CLT [14]:

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

4. see `https://en.wikipedia.org/wiki/Berry-Esseen_theorem`

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

$$\left| \mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0 \right) - \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma} \right) \right| \le \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}}. \tag{12}$$

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

$$\begin{aligned}
\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i \right] = \mathrm{sign}\,\mu \right) &= \mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0 \right) \\
&\overset{(12)}{\ge} \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma} \right) - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\mu\sqrt{\tau}/\sigma}^{\infty} e^{-t^2/2}\,dt - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\
&= \frac{1}{2}\left(1 + \sqrt{\frac{2}{\pi}}\int_{0}^{\mu\sqrt{\tau}/\sigma} e^{-t^2/2}\,dt - \frac{\nu^3}{\sigma^3\sqrt{\tau}} \right) \\
&= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma} \right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}} \right).
\end{aligned}$$

■

Clearly, bound (11) is better than (10) when randomness is high. On the other hand, bound (11) 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 10 *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*

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i \right] = \mathrm{sign}\,\mu \right) > \frac{1}{2}, \quad \textit{if} \quad \tau > 2\min\left(\frac{\sigma^2}{\mu^2}, \frac{\nu^3}{|\mu|\sigma^2} \right), \tag{13}$$

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 8. Assume both σ and ν are positive and finite.

In case of $\tau > 2\sigma^2/\mu^2$ we apply Lemma 8 again. Consider the case $\tau \le 2\sigma^2/\mu^2$, which implies $\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma} \le 1$. It is easy to check that $\mathrm{erf}(x)$ is concave on $[0,1]$ (in fact on $[0,\infty)$), therefore $\mathrm{erf}(x) \ge \mathrm{erf}(1)x$ for any $x \in [0,1]$. Setting $x = \frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma}$ we get

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

which together with (11) 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 (10) applying it to i.i.d. data $\hat{g}_i^1(x), \hat{g}_i^2(x), \ldots, \hat{g}_i^M(x)$.

C.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[5]. 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.

5. see https://en.wikipedia.org/wiki/Differential_entropy or https://en.wikipedia.org/wiki/Normal_distribution#Maximum_entropy

Remark 11 (see also Remark 12) *Note that SPB assumption describes the convergence of sign descent methods, which is known to be problem dependent (e.g. see [1], section 6.2 Results). One should view the SPB condition as a criteria to problems where sign based methods are useful.*

Remark 12 *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.*

C.5. Convergence Analysis: Proof of Theorem 2

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

$$\mathbb{E}[f(x_{k+1})|x_k] \le f(x_k) - \gamma_k \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] + \frac{d\bar{L}}{2}\gamma_k^2. \tag{14}$$

Using the definition of success probabilities ρ_i we get

$$\begin{aligned} \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] &= \langle g_k, \mathbb{E}[\operatorname{sign} \hat{g}_k] \rangle & (15) \\ &= \sum_{i=1}^{d} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] = \sum_{\substack{1 \le i \le d \\ g_{k,i} \ne 0}} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] & (16) \\ &= \sum_{\substack{1 \le i \le d \\ g_{k,i} \ne 0}} g_{k,i} \left(\rho_i(x_k) \operatorname{sign} g_{k,i} + (1 - \rho_i(x_k))(-\operatorname{sign} g_{k,i}) \right) & (17) \\ &= \sum_{\substack{1 \le i \le d \\ g_{k,i} \ne 0}} (2\rho_i(x_k) - 1)|g_{k,i}| = \sum_{i=1}^{d} (2\rho_i(x_k) - 1)|g_{k,i}| = \|g_k\|_\rho. & (18) \end{aligned}$$

Plugging this into (14) and taking full expectation, we get

$$\mathbb{E}\|g_k\|_\rho \le \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_k} + \frac{d\bar{L}}{2}\gamma_k. \tag{19}$$

Therefore

$$\sum_{k=0}^{K-1} \gamma_k \mathbb{E}\|g_k\|_\rho \le (f(x_0) - f^*) + \frac{d\bar{L}}{2} \sum_{k=0}^{K-1} \gamma_k^2. \tag{20}$$

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

$$\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \ge \sqrt{K}, \quad \sum_{k=1}^{K} \frac{1}{k} \le 2 + \log K. \tag{21}$$

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

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

C.6. Convergence Analysis: Proof of Theorem 13

Here we present slightly modified version of signSGD, where function evaluations are feasible. Algorithm 3 with Option 1 is the same Algorithm 1. We now state a general convergence rate for Algorithm 3 with Option 2.

Algorithm 3 SIGNSGD

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

Theorem 13 *Under the SPB assumption, Algorithm 3 (Option 2) with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as follows:*

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

In the case of constant step size $\gamma_k \equiv \gamma > 0$, the same rate as (6) is achieved.

Comparing Theorem 13 with Theorem 2, notice that a small modification in Algorithm 1 can remove the log-dependent factor from (5); we then 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 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.

Proof Clearly, the iterations $\{x_k\}_{k\ge 0}$ of Algorithm 3 (Option 2) do not increase the function value in any iteration, i.e. $\mathbb{E}[f(x_{k+1})|x_k] \le f(x_k)$. Continuing the proof of Theorem 2 from (19), we get

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

where we have used the following inequality

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

The proof for constant step size is the same as in Theorem 2. ■

C.7. Convergence Analysis in Distributed Setting: Proof of Theorem 4

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

$$I(p;a,b) = \frac{B(p;a,b)}{B(a,b)} = \frac{\int_0^p t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}, \quad a,b>0,\ p\in[0,1]. \tag{22}$$

The proof of Theorem 4 goes with the same steps as in Theorem 2, except the derivation (15)–(18) 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 14 *Assume that for some point $x \in \mathbb{R}^d$ and some coordinate $i \in \{1, 2, \ldots, d\}$, master node receives M independent stochastic signs $\operatorname{sign} \hat{g}_i^m(x)$, $m = 1, \ldots, M$ of true gradient $g_i(x) \neq 0$. Let $\hat{g}^{(M)}(x)$ be the sum of stochastic signs aggregated from nodes:*

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

Then

$$
\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(M)} \cdot g_i\right)\right] = 2I(\rho_i; l, l) - 1, \tag{23}
$$

where $l = \left[{}^{(M+1)}/{}_2\right]$ and $\rho_i > {}^1/{}_2$ is the success probablity for coordinate i.

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

$$
S_i^m := \begin{cases} 1, & \text{if } \operatorname{sign} \hat{g}_i^m = \operatorname{sign} g_i \\ 0, & \text{otherwise} \end{cases} \sim \text{Bernoulli}(\rho_i). \tag{24}
$$

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^m, which sum up to a binomial random variable S_i:

$$
S_i := \sum_{m=1}^{M} S_i^m \sim \text{Binomial}(M, \rho_i). \tag{25}
$$

First, let us consider the case when there are odd number of nodes, i.e. $M = 2l - 1$, $l \ge 1$. In this case, taking into account (24) and (25), we have

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

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

$$\mathrm{Prob}(S_i \ge l) = I(\rho_i; l, M-l+1) = I(\rho_i; l, l). \tag{26}$$

Therefore,

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

In the case of even number of nodes, i.e. $M = 2l$, $l \ge 1$, there is a probability to fail the vote $\mathrm{Prob}\left(\mathrm{sign}\, \hat{g}_i^{(M)} = 0\right) > 0$. However using (26) and properties of beta function[6] gives

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

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

C.8. Convergence Analysis in Distributed Setting: Variance reduction

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

Theorem 15 (Hoeffding's inequality for general bounded random variables; see [16], Theorem 2.2.6) *Let $X_1, X_2, \ldots, X_M$ be independent random variables. Assume that $X_m \in [A_m, B_m]$ for every m. Then, for any t ¿ 0, we have*

$$\mathrm{Prob}\left(\sum_{m=1}^M (X_m - \mathbb{E}X_m) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{m=1}^M (B_m - A_m)^2}\right).$$

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

$$X_i^m := \begin{cases} -1, & \text{if } \mathrm{sign}\, \hat{g}_i^m = \mathrm{sign}\, g_i \\ 1, & \text{otherwise} \end{cases} \tag{27}$$

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

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

6. see https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function

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

$$\mathrm{Prob}\left(\sum_{m=1}^{M} X_i^m \geq 0\right) \leq \exp\left(-\frac{1}{2}(2\rho_i - 1)^2 M\right).$$

Using Lemma 23, we get

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

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

Appendix D. Convergence Result for Standard SGD

For comparison, here we state and prove non-convex convergence rates of standard SGD with the same step sizes.

Theorem 16 (Non-convex convergence of SGD) *Let $\hat{g}$ be an unbiased estimator of the gradient ∇f and assume that $\mathbb{E}\|\hat{g}\|_2^2 \leq C$ for some $C > 0$. Then SGD with step sizes $\gamma_k = \gamma_0/\sqrt{k+1}$ converges as follows*

$$\min_{0 \leq k < K} \mathbb{E}\|\nabla f(x_k)\|_2^2 \leq \frac{1}{\sqrt{K}}\left[\frac{f(x_0) - f^*}{\gamma_0} + \gamma_0 C L_{\max}\right] + \frac{\gamma_0 C L_{\max}}{2}\frac{\log K}{\sqrt{K}}. \tag{28}$$

In the case of constant step size $\gamma_k \equiv \gamma > 0$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_2^2 \leq \frac{f(x_0) - f^*}{\gamma K} + \frac{C L_{\max}}{2}\gamma. \tag{29}$$

Proof From L-smoothness assumption we have

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})|x_k] &= \mathbb{E}[f(x_k - \gamma_k \hat{g}_k)|x_k] \\
&\leq f(x_k) - \mathbb{E}[\langle g_k, \gamma_k \hat{g}_k\rangle] + \frac{L_{\max}}{2}\gamma_k^2 \mathbb{E}[\|\hat{g}_k\|_2^2] \\
&= f(x_k) - \gamma_k \|g_k\|_2^2 + \frac{L_{\max}}{2}\gamma_k^2\, \mathbb{E}[\|\hat{g}_k\|_2^2].
\end{aligned}$$

Taking full expectation, using variance bound assumption, we have

$$\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(x_k)] \leq -\gamma_k\, \mathbb{E}\|g_k\|_2^2 + \frac{L_{\max}}{2}\gamma_k^2 C$$

Therefore

$$\gamma_k \mathbb{E}\|g_k\|_2^2 \leq \mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})] + \frac{C L_{\max}}{2}\gamma_k^2$$

Summing $k = 0, 1, \dots, K-1$ gives

$$\sum_{k=0}^{K-1} \gamma_k \mathbb{E}\|g_k\|_2^2 \le (f(x_0) - f^*) + \frac{CL_{\max}}{2} \sum_{k=0}^{K-1} \gamma_k^2.$$

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

$$\begin{aligned}
\min_{0\le k<K} \mathbb{E}\|g_k\|_2^2 &\le \sum_{k=0}^{K-1} \frac{\gamma_0}{\sqrt{k+1}} \mathbb{E}\|g_k\|_2^2 \Big/ \sum_{k=0}^{K-1} \frac{\gamma_0}{\sqrt{k+1}} \\
&\le \frac{1}{\sqrt{K}} \left[\frac{f(x_0)-f^*}{\gamma_0} + \frac{CL_{\max}}{2}\gamma_0 \sum_{k=0}^{K-1} \frac{1}{k+1} \right] \\
&\le \frac{1}{\sqrt{K}} \left[\frac{f(x_0)-f^*}{\gamma_0} + \gamma_0 CL_{\max} + \frac{\gamma_0 CL_{\max}}{2} \log K \right] \\
&= \frac{1}{\sqrt{K}} \left[\frac{f(x_0)-f^*}{\gamma_0} + \gamma_0 CL_{\max} \right] + \frac{\gamma_0 CL_{\max}}{2} \frac{\log K}{\sqrt{K}}.
\end{aligned}$$

where again we have used inequalities (21). In the case of constant step size $\gamma_k = \gamma$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\|_2^2 \le \frac{1}{\gamma K}\left[(f(x_0)-f^*) + \frac{CL_{\max}}{2}\gamma^2 K \right] = \frac{f(x_0)-f^*}{\gamma K} + \frac{CL_{\max}}{2}\gamma.$$

■

Appendix E. Recovering Theorem 1 in [3] from Theorem 2

To recover Theorem 1 in [3], first note that choosing a particular step size γ in (6) yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\|_\rho \le \sqrt{\frac{2d\bar{L}(f(x_0)-f^*)}{K}}, \quad \text{with} \quad \gamma = \sqrt{\frac{2(f(x_0)-f^*)}{d\bar{L}K}}. \tag{30}$$

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 [3]: let $H_k = \{1 \le i \le d \colon \sigma_i < \sqrt{3}/2|g_{k,i}|\}$

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

Appendix F. Stochastic signSGD

Our experiments and the counterexample show that signSGD might fail to converge in general. What we proved is that SPB assumption is roughly a necessary and sufficient for general convergence. There are several ways to overcome SPB assumption and make signSGD to work in general, e.g. scaled version of signSGD with error feedback [7]. Here we to present a simple way of fixing this issue, which is more natural to signSGD. The issue with signSGD is that sign of stochastic gradient is biased, which also complicates the analysis.

We define stochastic sign operator $\widetilde{\mathrm{sign}}$, which unlike the deterministic sign operator is unbiased with appropriate scaling factor.

Definition 17 (Stochastic Sign) *Define the stochastic sign operator* $\widetilde{\mathrm{sign}} : \mathbb{R}^d \to \mathbb{R}^d$ *as*

$$\left(\widetilde{\mathrm{sign}}\, g\right)_i = \begin{cases} +1, & \textit{with prob. } \frac{1}{2} + \frac{1}{2}\frac{g_i}{\|g\|_2} \\ -1, & \textit{with prob. } \frac{1}{2} - \frac{1}{2}\frac{g_i}{\|g\|_2} \end{cases}, \quad 1 \le i \le d,$$

and $\widetilde{\mathrm{sign}}\, \mathbf{0} = \mathbf{0}$ *with probability 1.*

Furthermore, we define stochastic compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ as $\mathcal{C}(x) = \|x\|_2 \cdot \widetilde{\mathrm{sign}}\, x$, which compresses rd bits to $r + d$ bits (r bits per one floating point number). Then for any unbiased estimator $\hat{g}$ we get

$$\mathbb{E}\left[\mathcal{C}(\hat{g})\right] = \mathbb{E}\left[\mathbb{E}[\mathcal{C}(\hat{g}) \mid \hat{g}]\right] = \mathbb{E}\left[\|\hat{g}\|_2 \left(\frac{1}{2} + \frac{1}{2}\frac{\hat{g}}{\|\hat{g}\|_2}\right) - \|\hat{g}\|_2 \left(\frac{1}{2} - \frac{1}{2}\frac{\hat{g}}{\|\hat{g}\|_2}\right)\right] = \mathbb{E}[\hat{g}] = g,$$

$$\mathrm{Var}\left[\mathcal{C}(\hat{g})\right] = \mathbb{E}\left[\|\mathcal{C}(\hat{g}) - \hat{g}\|_2^2\right] = \mathbb{E}\left[\|\mathcal{C}(\hat{g})\|_2^2\right] - \mathbb{E}\left[\|\hat{g}\|_2^2\right] = (d-1)\mathbb{E}\|\hat{g}\|_2^2.$$

Using this relations, any analysis for SGD can be repeated for stochastic signSGD giving the same convergence rate with less communication and with $(d-1)$ times worse coefficients.

Another scaled version of signSGD investigated in [7] uses non-stochastic compression operator $\mathcal{C}' : \mathbb{R}^d \to \mathbb{R}^d$ defined as $\mathcal{C}'(x) = \frac{\|x\|_1}{d}\, \mathrm{sign}\, x$. It is shown (see [7], Theorem II) to converge as

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_2^2 \le \frac{2\left(f(x_0) - f^*\right)}{\gamma K} + \frac{\gamma L_{\max} C}{2} + 4d(d-1)\gamma^2 L_{\max}^2 C,$$

where the error of current gradient compression is stored to be used in the next step. On the other hand, adopting the analysis of Theorem 16 for the stochastic compression operator $\mathcal{C}$, we get a bound

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|_2^2 \le \frac{f(x_0) - f^*}{\gamma K} + \frac{\gamma L_{\max} C d}{2},$$

where no data needs to be stored. Furthermore, ignoring the factor 2 at the first term, later bound is better if $\gamma \ge 1/8dL_{\max}$.