Appendix: “Stochastic Sign Descent Methods: New Algorithms and Better Theory”

A. Additional Experiments

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

Figure 5 shows the robustness of SPB assumption in the convergence rate (10) 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.

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.

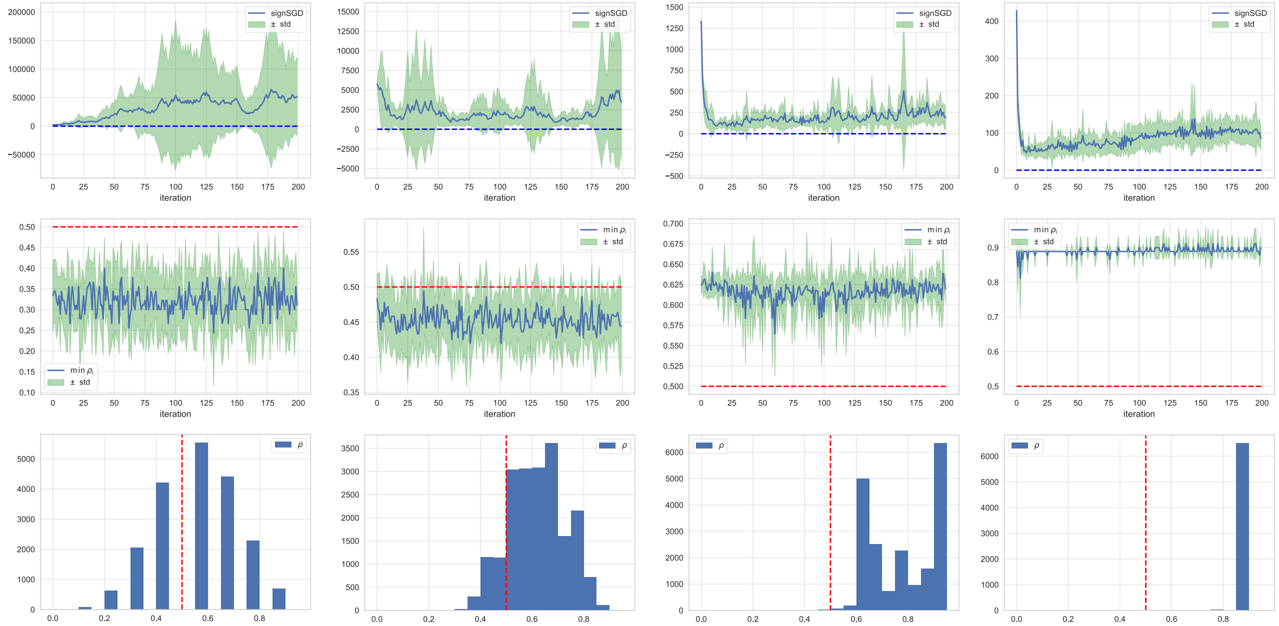

Figure 5. Performance of signSGD with constant step size ($\gamma = 0.25$) under four different noise levels (mini-batch size 1, 2, 5, 8) using Rosenbrock function. Each column represent a separate experiment with function values, evolution of minimum success probabilities and the histogram of success 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.

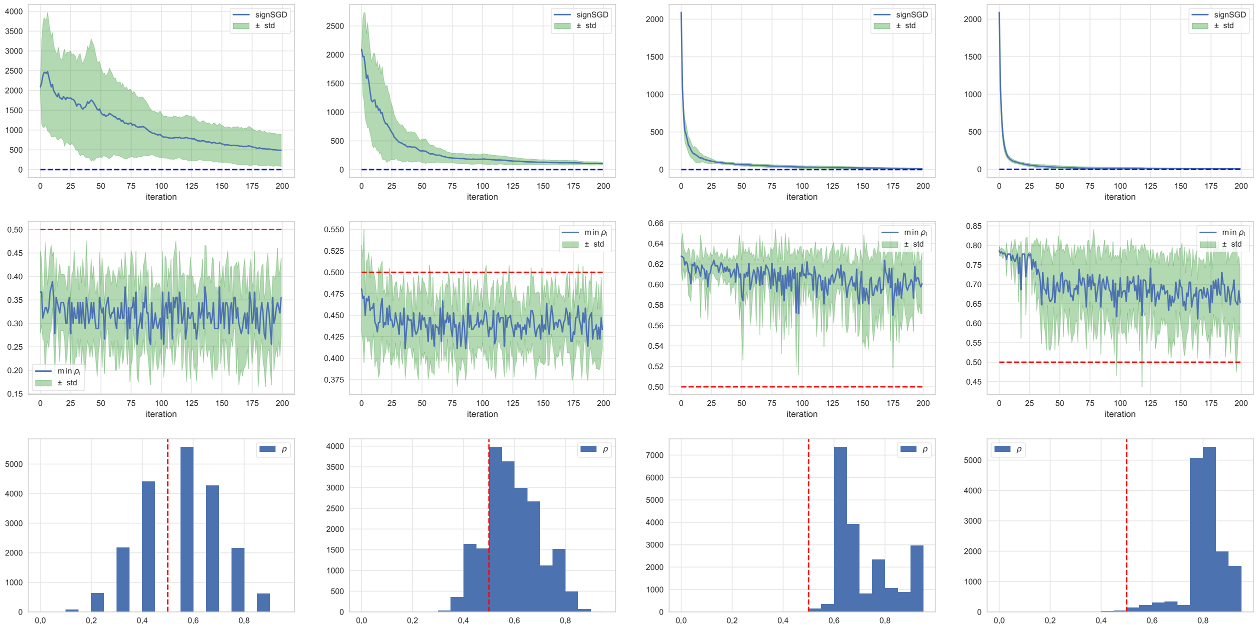

Figure 6. Performance of signSGD with variable step size ($\gamma_0 = 0.25$) under four different noise levels (mini-batch size 1, 2, 5, 7) using Rosenbrock function. As in the experiments of Figure 5 with constant step size, these plots show the relationship between success probabilities and the convergence rate (9). 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.

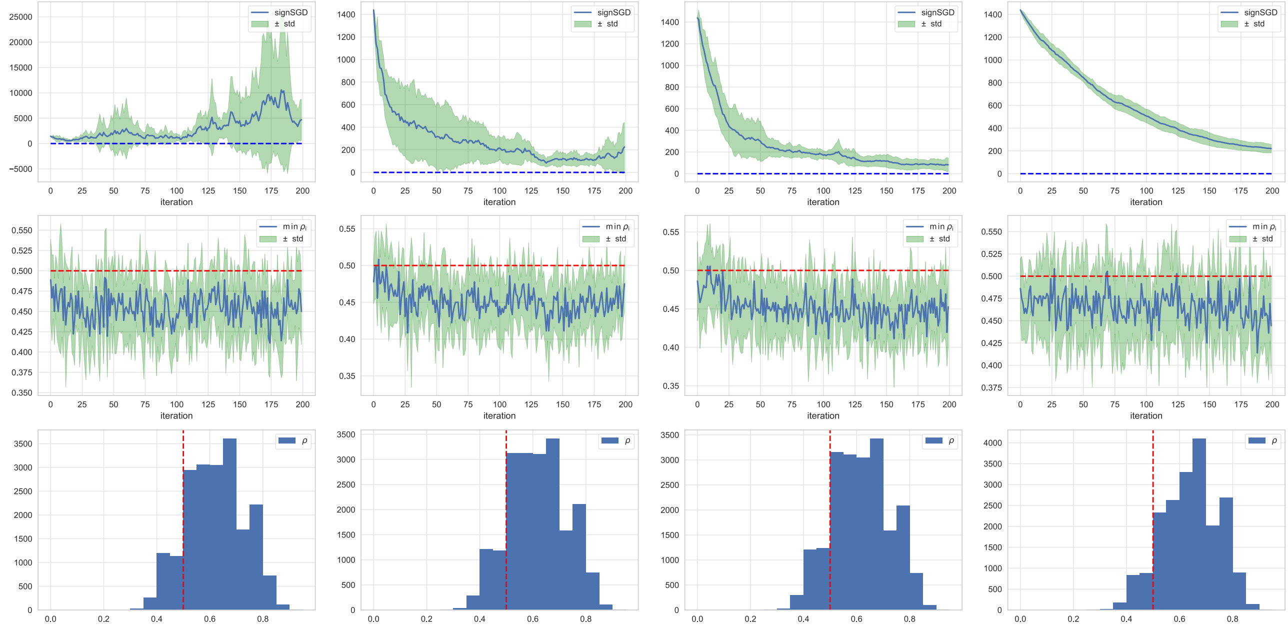

Figure 7. In this part of experiments we investigated convergence rate (10) to a neighborhood of the solution. We fixed gradient noise level by setting mini-batch size 2 and altered the constant step size. For the first column we set bigger step size $\gamma = 0.25$ to detect the divergence (as we slightly 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.

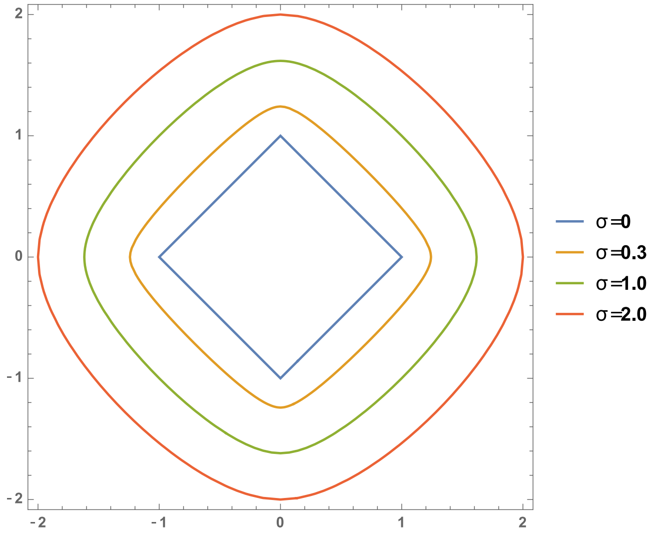

Figure 8. Unit balls in $l^{1,2}$ norm (7) with different noise levels.

B. More details on ρ-norm and $l^{1,2}$ norm

In this section we give more details on the concept of a norm-like function, which we call *ρ-norm*. First, we recall the definition from the main part of the paper.

Definition 4 (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 1, ρ-norm can be lower bounded by a mixture of the l^1 and squared l^2 norms:

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

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.

C. Proofs

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

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

Theorem 5 (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,*

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

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

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

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 $X = \hat{g}_i(x)$, $\mu =$

---

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

$g_i(x), \sigma^2 = \sigma_i^2(x)$ and get a bound for success probabilities

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

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

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

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

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

Hence, continuing the proof of Lemma 1, we get

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

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 $\operatorname{sign}\hat{g}^{(\tau)}$ when $g_i \neq 0$, we have

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

which imples

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

C.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 5. *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 $\tau \geq 1$*

$$\operatorname{Prob}\left(\operatorname{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \operatorname{sign}\mu\right) \geq 1 - \frac{\sigma^2}{\tau\mu^2}. \tag{13}$$

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

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

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

Obviously, bound (13) 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 6. *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 $\tau \geq 1$*

$$\operatorname{Prob}\left(\operatorname{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \operatorname{sign}\mu\right) \geq \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{|\mu|\sqrt{\tau}}{\sqrt{2}\sigma}\right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right), \tag{14}$$

where error function erf *is defined as*

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

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):

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

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

$$\left| \mathrm{Prob}\left(\frac{1}{\tau} \sum_{i=1}^{\tau} X_i > 0 \right) - \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma}\right) \right| \le \frac{1}{2} \frac{\nu^3}{\sigma^3\sqrt{\tau}}. \tag{15}$$

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

$$\begin{aligned}
\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu \right) &= \mathrm{Prob}\left(\frac{1}{\tau}\sum_{i=1}^{\tau} X_i > 0 \right) \\
&\overset{(15)}{\ge} \mathrm{Prob}\left(N > -\frac{\mu\sqrt{\tau}}{\sigma}\right) - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\mu\sqrt{\tau}/\sigma}^{\infty} e^{-t^2/2}\,dt - \frac{1}{2}\frac{\nu^3}{\sigma^3\sqrt{\tau}} \\
&= \frac{1}{2}\left(1 + \sqrt{\frac{2}{\pi}} \int_0^{\mu\sqrt{\tau}/\sigma} e^{-t^2/2}\,dt - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right) \\
&= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mu\sqrt{\tau}}{\sqrt{2}\sigma}\right) - \frac{\nu^3}{\sigma^3\sqrt{\tau}}\right).
\end{aligned}$$

□

Clearly, bound (14) is better than (13) when randomness is high. On the other hand, bound (14) 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 7. *Let $X_1, X_2, \ldots, X_\tau$ be i.i.d. random variables with non-zero mean $\mu \ne 0$ and finite variance $\sigma^2 < \infty$. Then*

$$\mathrm{Prob}\left(\mathrm{sign}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right] = \mathrm{sign}\,\mu \right) > \frac{1}{2}, \quad \textit{if} \quad \tau > 2\min\left(\frac{\sigma^2}{\mu^2}, \frac{\nu^3}{|\mu|\sigma^2}\right), \tag{16}$$

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

In case of $\tau > 2\sigma^2/\mu^2$ we apply Lemma 5 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

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

---

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

which together with (14) gives

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

Hence, SPB assumption holds if

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

It remains to show that $\operatorname{erf}(1)>{}^1\!/\!\sqrt{2}$. 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

$$\begin{aligned}\operatorname{erf}(1)&=\frac{2}{\sqrt{\pi}}\int_0^1 e^{-t^2}\,dt\\&\geq\frac{2}{\sqrt{\pi}}\int_0^1\left(1-(1-{}^1\!/\!e)t^2\right)\,dt\\&=\frac{2}{\sqrt{\pi}}\left(\frac{2}{3}+\frac{1}{3e}\right)>\frac{2}{\sqrt{4}}\left(\frac{2}{3}+\frac{1}{3\cdot3}\right)=\frac{7}{9}>\frac{1}{\sqrt{2}}.\end{aligned}$$

□

Lemma (3) follows from Lemma (7) 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[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 $\operatorname{Prob}(\operatorname{sign}\hat{g}=\operatorname{sign}g)>{}^1\!/\!2$ 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 $g\neq0$. 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.

C.5. Convergence Analysis for $M=1$: Proof of Theorem 1

Basically, the analysis follows the standard steps used to analyze SGD for non-convex objectives, except the part (18)–(21) where inner product $\mathbb{E}[\langle g_k,\operatorname{sign}\hat{g}_k\rangle]$ needs to be estimated. This is exactly the place when stochastic gradient estimator $\operatorname{sign}\hat{g}_k$ 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

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

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

$$\mathbb{E}[f(x_{k+1}) | x_k] \le f(x_k) - \gamma_k \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] + \frac{d\bar{L}}{2} \gamma_k^2. \tag{17}$$

Using the definition of success probabilities ρ_i we get

$$\begin{aligned} \mathbb{E}[\langle g_k, \operatorname{sign} \hat{g}_k \rangle] &= \langle g_k, \mathbb{E}[\operatorname{sign} \hat{g}_k] \rangle & (18) \\ &= \sum_{i=1}^{d} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] = \sum_{\substack{1 \le i \le d \\ g_{k,i} \neq 0}} g_{k,i} \cdot \mathbb{E}[\operatorname{sign} \hat{g}_{k,i}] & (19) \\ &= \sum_{\substack{1 \le i \le d \\ g_{k,i} \neq 0}} g_{k,i} \left(\rho_i(x_k) \operatorname{sign} g_{k,i} + (1 - \rho_i(x_k))(-\operatorname{sign} g_{k,i}) \right) & (20) \\ &= \sum_{\substack{1 \le i \le d \\ g_{k,i} \neq 0}} (2\rho_i(x_k) - 1)|g_{k,i}| = \sum_{i=1}^{d} (2\rho_i(x_k) - 1)|g_{k,i}| = \|g_k\|_\rho. & (21) \end{aligned}$$

Plugging this into (17) and taking full expectation, we get

$$\mathbb{E}\|g_k\|_\rho \le \frac{\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]}{\gamma_k} + \frac{d\bar{L}}{2} \gamma_k. \tag{22}$$

Therefore

$$\sum_{k=0}^{K-1} \gamma_k \mathbb{E}\|g_k\|_\rho \le (f(x_0) - f^*) + \frac{d\bar{L}}{2} \sum_{k=0}^{K-1} \gamma_k^2. \tag{23}$$

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

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

where we have used the following standard inequalities

$$\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \ge \sqrt{K}, \quad \sum_{k=1}^{K} \frac{1}{k} \le 2 + \log K. \tag{24}$$

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

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

Thus, we complete the proof of Theorem 1. Next, we consider setups of Lemma 1 and Lemma 2 to simplify convergence rates (9) and (10) of signSGD without the generic ρ-norm. Combining Lemma 1 and Theorem 1, we get

Corollary 1. *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 $g(x)=\nabla f(x)$. If for each coordinate $\hat{g}_i$ has a unimodal and symmetric distribution with variance $\sigma_i^2=\sigma_i^2(x)$, $1\le i\le d$, then single node signSGD (Algorithm 1) with Option 1 and with step sizes $\gamma_k=\gamma_0/\sqrt{k+1}$ converges as follows*

$$\min_{0\le k<K}\mathbb{E}\left[\sum_{i=1}^d \frac{|g_i(x_k)|^2}{|g_i(x_k)|+\sqrt{3}\sigma_i(x_k)}\right] \le \frac{f(x_0)-f^*}{\gamma_0\sqrt{K}} + \frac{3\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}\ . \tag{25}$$

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

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sum_{i=1}^d \frac{|g_i(x_k)|^2}{|g_i(x_k)|+\sqrt{3}\sigma_i(x_k)}\right] \le \frac{f(x_0)-f^*}{\gamma K} + \frac{\gamma d\bar{L}}{2}\ . \tag{26}$$

As we discussed in Section 3.3 of the main part, in this case gradient appears through a mixture of l^1 and squared l^2 norms (7). Similar mixed norm convergence rates for signSGD was established by Bernstein et al. (2019) (see Theorem 1) and by Chen et al. (2020) (see Theorem 5). Furthermore, combining Lemma 2 and Theorem 1, we get

Corollary 2. *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 $g(x)=\nabla f(x)$ and coordinate-wise variances $\sigma_i^2(x)\le c_i\, g_i^2(x)$ are bounded for some constants c_i. If mini-batch size $\tau>2\max_{1\le i\le d}c_i$ then single node signSGD (Algorithm 1) with Option 1 and with step sizes $\gamma_k=\gamma_0/\sqrt{k+1}$ converges as follows*

$$\min_{0\le k<K}\mathbb{E}\left[\sum_{i=1}^d w_i|g_i(x_k)|\right] \le \frac{f(x_0)-f^*}{\gamma_0\sqrt{K}} + \frac{3\gamma_0 d\bar{L}}{2}\frac{\log K}{\sqrt{K}}, \tag{27}$$

where $w_i = 1-\frac{2c_i}{\tau}$ are positive weights. If $\gamma_k\equiv\gamma>0$, we get $1/K$ convergence to a neighbourhood:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sum_{i=1}^d w_i|g_i(x_k)|\right] \le \frac{f(x_0)-f^*}{\gamma K} + \frac{\gamma d\bar{L}}{2}\ . \tag{28}$$

Notice that in this case gradient appears through weighted l^1 norm (8).

C.6. Convergence Analysis for $M=1$: Proof of Theorem 2

Clearly, the iterations $\{x_k\}_{k\ge0}$ of Algorithm 1 with Option 2 do not increase the function value in any iteration, i.e. $\mathbb{E}[f(x_{k+1})|x_k]\le f(x_k)$. Continuing the proof of Theorem 1 from (22), 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 1.

C.7. Convergence Analysis in Parallel Setting: Proof of Theorem 3

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

$$I(p;a,b) = \frac{B(p;a,b)}{B(a,b)} = \frac{\int_0^p t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}, \quad a,b>0,\; p\in[0,1]. \tag{29}$$

The proof of Theorem 3 goes with the same steps as in Theorem 1, except the derivation (18)–(21) 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 8. *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 $\operatorname{sign}\hat{g}_i^m(x)$, $m=1,\dots,M$ of true gradient $g_i(x)\neq 0$. Let $\hat{g}^{(M)}(x)$ be the sum of stochastic signs aggregated from nodes:*

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

Then

$$\mathbb{E}\left[\operatorname{sign}\left(\hat{g}_i^{(M)}\cdot g_i\right)\right] = 2I(\rho_i;l,l)-1, \tag{30}$$

where $l = [^{(M+1)}/_2]$ 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:

$$S_i^m := \begin{cases} 1, & \text{if } \operatorname{sign}\hat{g}_i^m = \operatorname{sign} g_i \\ 0, & \text{otherwise} \end{cases} \sim \text{Bernoulli}(\rho_i). \tag{31}$$

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:

$$S_i := \sum_{m=1}^{M} S_i^m \sim \text{Binomial}(M,\rho_i). \tag{32}$$

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 (31) and (32), 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:

$$\mathrm{Prob}(S_i \ge l) = I(\rho_i; l, M-l+1) = I(\rho_i; l, l). \tag{33}$$

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 (33) and properties of beta function[5] 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 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 6 (Hoeffding's inequality for general bounded random variables; see (Vershynin, 2018), Theorem 2.2.6)**.** *Let $X_1, X_2, \dots, X_M$ 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, $m = 1, 2, \dots, M$ showing the missmatch between stochastic gradient sign and full gradient sign from node m and coordinate i:

$$X_i^m := \begin{cases} -1, & \text{if } \mathrm{sign}\,\hat{g}_i^m = \mathrm{sign}\,g_i \\ 1, & \text{otherwise} \end{cases} \tag{34}$$

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

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

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

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

Using Lemma 30, we get

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

---

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

which provides the following estimate for ρ_M-norm:

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

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

C.9. Distributed Training with Partitioned Data: Proof of Theorem 4

We follow the analysis of Cutkosky & 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 9 (see Lemma 2 in (Cutkosky & Mehta, 2020))**.** *For any non-zero vectors a and b*

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

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

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

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

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

□

We start from the smoothness of functions f_n

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

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

$$\begin{aligned}
\mathbb{E}\left[f(x_{k+1})|x_k,\hat{g}_k^n\right] &\le f(x_k) - \frac{\gamma}{M}\sum_{n=1}^{M}\frac{\langle g_k, m_k^n\rangle}{\|m_k^n\|} + \frac{d\tilde{L}\gamma^2}{2}\\
&\overset{\text{Lemma 9}}{\le} f(x_k) - \frac{\gamma}{3}\|g_k\| + \frac{8\gamma}{3M}\sum_{n=1}^{M}\|m_k^n - g_k\| + \frac{d\tilde{L}\gamma^2}{2}.
\end{aligned} \tag{35}$$

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

$$\mathbb{E}\left[\langle \epsilon_k^n, \epsilon_{k'}^n\rangle\right] \begin{cases} \le \sigma_n^2 & \text{if } k = k', \\ = 0 & \text{if } k \ne k'. \end{cases} \tag{36}$$

Using L_n-smoothness of functions f_n again, we have

$$\|g_k - g_{k+1}\| \le \frac{1}{M}\sum_{n=1}^{M}\|g_k^n - g_{k+1}^n\| \le \frac{1}{M}\sum_{n=1}^{M} L_n\|x_k - x_{k+1}\| = \frac{\tilde{L}\gamma}{M}\|s_k\| \le \tilde{L}\gamma\sqrt{d}. \tag{37}$$

Therefore

$$\begin{aligned}
\mathbb{E}\|\hat{\epsilon}_{k+1}^n\| &\le \beta^{k+1}\|\epsilon_0^n\| + \sum_{t=0}^{k}\beta^{t+1}\|g_{k-t} - g_{k+1-t}\| + (1-\beta)\mathbb{E}\left\|\sum_{t=0}^{k}\beta^t\epsilon_{k+1-t}^n\right\| \\
&\overset{(37)}{\le} \beta^{k+1}\sigma_n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + (1-\beta)\sqrt{\mathbb{E}\left\|\sum_{t=0}^{k}\beta^t\epsilon_{k+1-t}^n\right\|^2} \\
&\overset{(36)}{\le} \beta^{k+1}\sigma_n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + (1-\beta)\sqrt{\sum_{t=0}^{k}\beta^{2t}\sigma_n^2} \\
&\le \beta^{k+1}\sigma_n + \frac{\tilde{L}\gamma\sqrt{d}}{1-\beta} + \sigma_n\sqrt{1-\beta}
\end{aligned}$$

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 (35), 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

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

D. Recovering Theorem 1 in (Bernstein et al., 2019) from Theorem 1

To recover Theorem 1 in (Bernstein et al., 2019), first note that choosing a particular step size γ in (10) yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\|_\rho \le \sqrt{\frac{2d\bar{L}(f(x_0)-f^*)}{K}}, \quad \text{with} \quad \gamma = \sqrt{\frac{2(f(x_0)-f^*)}{d\bar{L}K}}. \tag{38}$$

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

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