A. Further experimental results

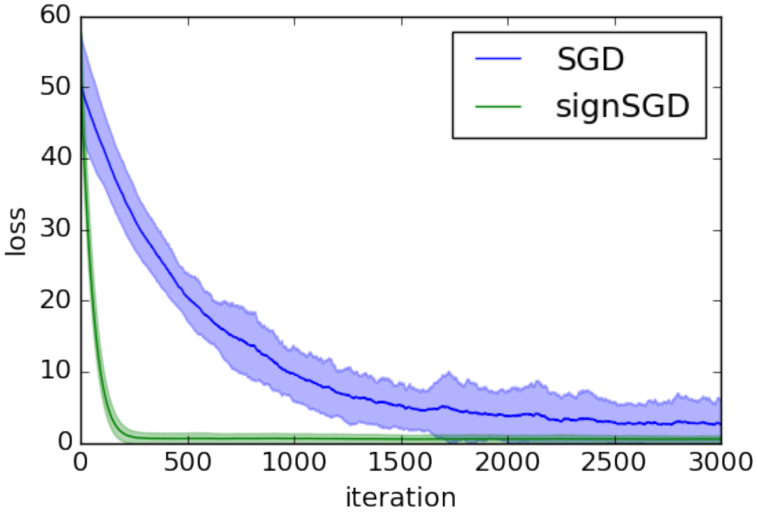

Figure A.1. A simple toy problem where SIGNSGD converges faster than SGD. The objective function is just a quadratic $f(x) = \frac{1}{2}x^2$ for $x \in \mathbb{R}^{100}$. The gradient of this function is just $g(x) = x$. We construct an artificial stochastic gradient by adding Gaussian noise $\mathcal{N}(0, 100^2)$ to only the first component of the gradient. Therefore the noise is extremely sparse. The initial point is sampled from a unit variance spherical Gaussian. For each algorithm we tune a separate, constant learning rate finding 0.001 best for SGD and 0.01 best for SIGNSGD. SIGNSGD appears more robust to the sparse noise in this problem. Results are averaged over 50 repeats with ±1 standard deviation shaded.

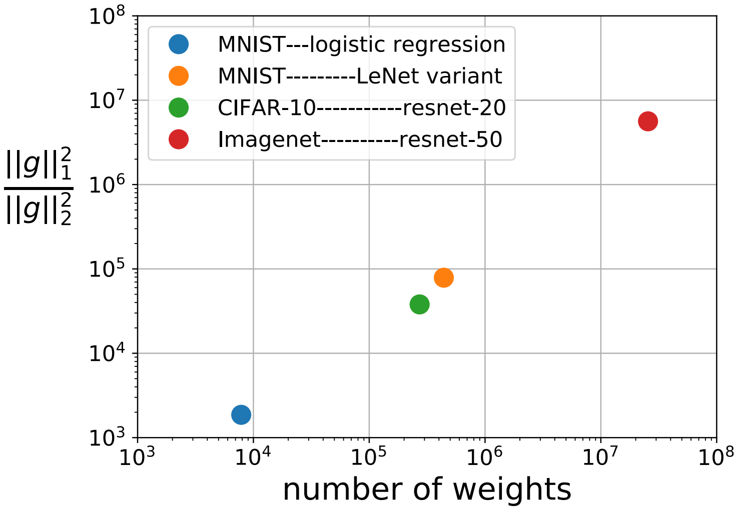

Figure A.2. Measuring gradient density via ratio of norms, over a range of datasets and architectures. For each network, we take a point in parameter space provided by the Xavier initialiser (Glorot & Bengio, 2010). We do a full pass over the data to compute the full gradient at this point. It is remarkably dense in all cases.

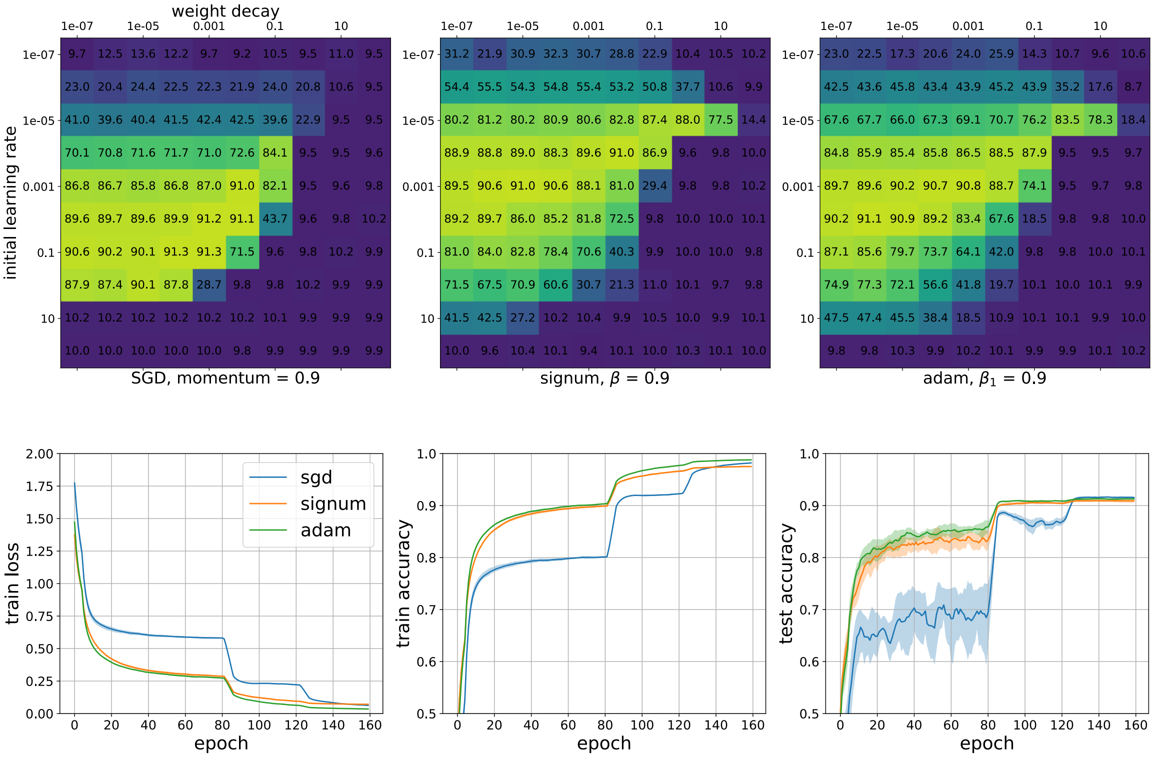

Figure A.3. CIFAR-10 results using SIGNUM to train a Resnet-20 model. Top: validation accuracies from a hyperparameter sweep on a separate validation set carved out from the training set. We used this to tune initial learning rate, weight decay and momentum for all algorithms. All other hyperparameter settings were chosen as in (He et al., 2016a) as found favourable for SGD. The hyperparameter sweep for other values of momentum is plotted in Figure A.4 of the supplementary. Bottom: there is little difference between the final test set performance of the algorithms. SIGNUM closely resembles ADAM in all of these plots.

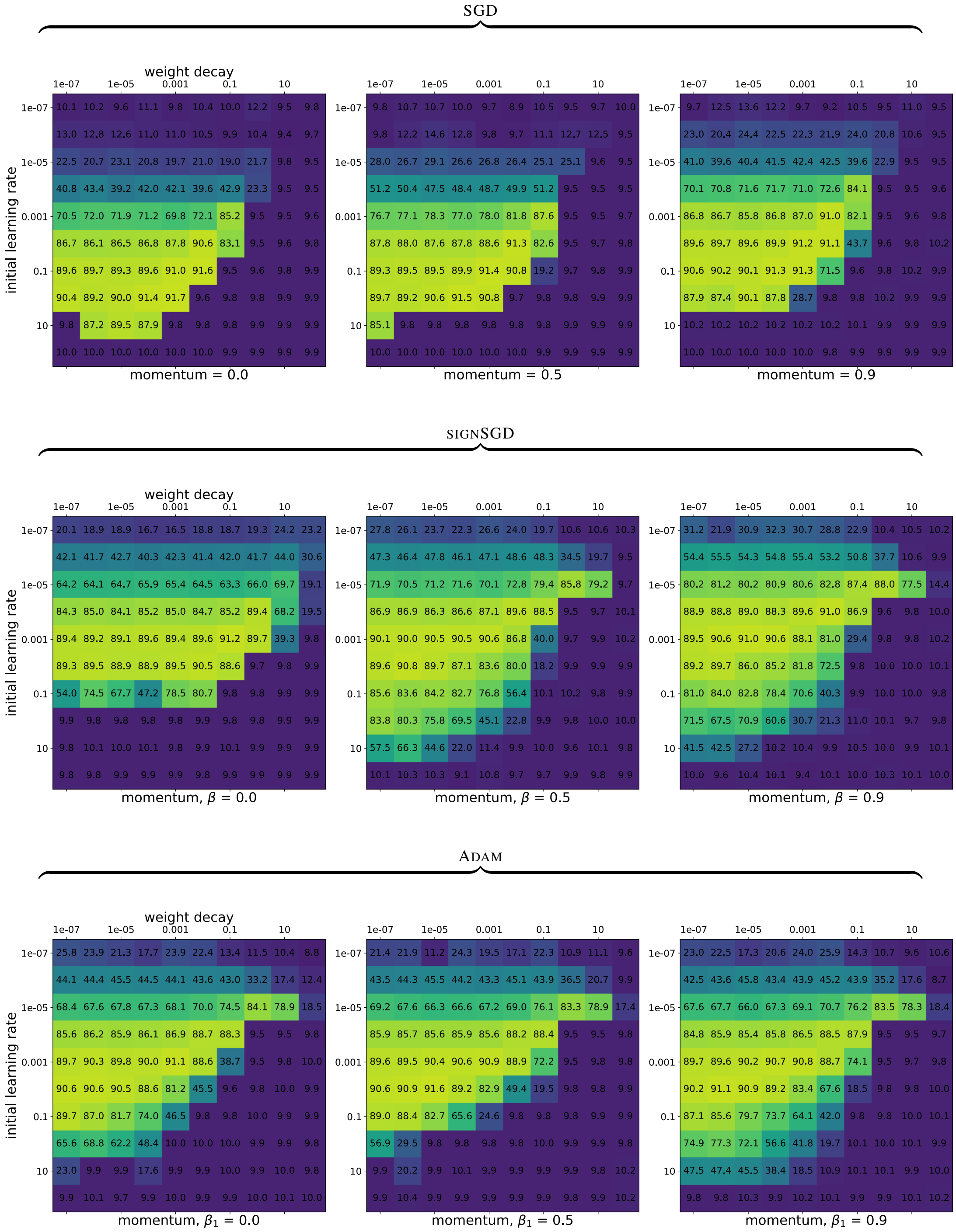

Figure A.4. Results of a massive grid search over hyperparameters for training Resnet-20 (He et al., 2016a) on CIFAR-10 (Krizhevsky, 2009). All non-algorithm specific hyperparameters (such as learning rate schedules) were set as in (He et al., 2016a). In ADAM, β_2 and ϵ were chosen as recommended in (Kingma & Ba, 2015). Data was divided according to a $\{45k/5k/10k\}$ $\{train/val/test\}$ split. Validation accuracies are plotted above, and the best performer on the validation set was chosen for the final test run (shown in Figure A.3). All algorithms at the least get close to the baseline reported in (He et al., 2016a) of 91.25%. Note the broad similarity in general shape of the heatmap between ADAM and SIGNSGD, supporting a notion of algorithmic similarity. Also note that whilst SGD has a larger region of very high-scoring hyperparameter configurations, SIGNSGD and ADAM appear stable over a larger range of learning rates.

B. Proving the convergence rate of SIGNSGD

Theorem 1 (Non-convex convergence rate of SIGNSGD)**.** *Run algorithm 1 for K iterations under Assumptions 1 to 3. Set the learning rate and mini-batch size (independently of step k) as*

$$\delta_k = \frac{1}{\sqrt{\|\vec{L}\|_1 K}}, \qquad n_k = K$$

Let N be the cumulative number of stochastic gradient calls up to step K, i.e. $N = O(K^2)$. Then we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_1\right]^2 \leq \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + 2\|\vec{\sigma}\|_1\right]^2$$

Proof. First let's bound the improvement of the objective during a single step of the algorithm for one instantiation of the noise. $\mathbb{I}[.]$ is the indicator function, $g_{k,i}$ denotes the i^{th} component of the true gradient $g(x_k)$ and $\tilde{g}_k$ is a stochastic sample obeying Assumption 3.

First take Assumption 2, plug in the step from Algorithm 1, and decompose the improvement to expose the stochasticity-induced error:

$$\begin{aligned} f_{k+1} - f_k &\leq g_k^T(x_{k+1} - x_k) + \sum_{i=1}^{d}\frac{L_i}{2}(x_{k+1} - x_k)_i^2 \\ &= -\delta_k g_k^T \operatorname{sign}(\tilde{g}_k) + \delta_k^2 \sum_{i=1}^{d}\frac{L_i}{2} \\ &= -\delta_k\|g_k\|_1 + \frac{\delta_k^2}{2}\|\vec{L}\|_1 \\ &\quad + 2\delta_k\sum_{i=1}^{d}|g_{k,i}|\,\mathbb{I}[\operatorname{sign}(\tilde{g}_{k,i}) \neq \operatorname{sign}(g_{k,i})] \end{aligned}$$

Next we find the expected improvement at time $k+1$ conditioned on the previous iterate.

$$\begin{aligned} \mathbb{E}[f_{k+1} - f_k|x_k] \leq -\delta_k\|g_k\|_1 + \frac{\delta_k^2}{2}\|\vec{L}\|_1 \\ + 2\delta_k\sum_{i=1}^{d}|g_{k,i}|\,\mathbb{P}[\operatorname{sign}(\tilde{g}_{k,i}) \neq \operatorname{sign}(g_{k,i})] \end{aligned}$$

So the expected improvement crucially depends on the probability that each component of the sign vector is correct, which is intuitively controlled by the relative scale of the gradient to the noise. To make this rigorous, first relax the probability, then use Markov's inequality followed by Jensen's inequality:

$$\begin{aligned} \mathbb{P}[\operatorname{sign}(\tilde{g}_{k,i}) \neq \operatorname{sign}(g_{k,i})] &\leq \mathbb{P}[|\tilde{g}_{k,i} - g_{k,i}| \geq |g_{k,i}|] \\ &\leq \frac{\mathbb{E}[|\tilde{g}_{k,i} - g_{k,i}|]}{|g_{k,i}|} \\ &\leq \frac{\sqrt{\mathbb{E}[(\tilde{g}_{k,i} - g_{k,i})^2]}}{|g_{k,i}|} \\ &= \frac{\sigma_{k,i}}{|g_{k,i}|} \end{aligned}$$

$\sigma_{k,i}$ refers to the variance of the k^{th} stochastic gradient estimate, computed over a mini-batch of size n_k. Therefore, by Assumption 3, we have that $\sigma_{k,i} \leq \sigma_i / \sqrt{n_k}$.

We now substitute these results and our learning rate and mini-batch settings into the expected improvement:

$$\begin{aligned}\mathbb{E}[f_{k+1} - f_k | x_k] &\leq -\delta_k \|g_k\|_1 + 2\frac{\delta_k}{\sqrt{n_k}}\|\vec{\sigma}\|_1 + \frac{\delta_k^2}{2}\|\vec{L}\|_1 \\ &= -\frac{1}{\sqrt{\|\vec{L}\|_1 K}}\|g_k\|_1 + \frac{2}{\sqrt{\|\vec{L}\|_1 K}}\|\vec{\sigma}\|_1 + \frac{1}{2K}\end{aligned}$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the iterations:

$$\begin{aligned}f_0 - f^* &\geq f_0 - \mathbb{E}[f_K] \\ &= \mathbb{E}\left[\sum_{k=0}^{K-1} f_k - f_{k+1}\right] \\ &\geq \mathbb{E}\sum_{k=0}^{K-1}\left[\frac{1}{\sqrt{\|\vec{L}\|_1 K}}\|g_k\|_1 - \frac{1}{2\sqrt{\|\vec{L}\|_1 K}}\left(4\|\sigma\|_1 + \sqrt{\|\vec{L}\|_1}\right)\right] \\ &= \sqrt{\frac{K}{\|\vec{L}\|_1}}\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_1\right] - \frac{1}{2\sqrt{\|\vec{L}\|_1}}\left(4\|\vec{\sigma}\|_1 + \sqrt{\|\vec{L}\|_1}\right)\end{aligned}$$

We can rearrange this inequality to yield the rate:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_1\right] \leq \frac{1}{\sqrt{K}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + 2\|\vec{\sigma}\|_1\right]$$

Since we are growing our mini-batch size, it will take $N = O(K^2)$ gradient calls to reach step K. Substitute this in, square the result, and we are done. □

C. Large and small batch SGD

Algorithm C.1 SGD

```
  **Input:** learning rate $\delta$, current point $x_k$
  $\tilde{g}_k \leftarrow \text{stochasticGradient}(x_k)$
  $x_{k+1} \leftarrow x_k - \delta\,\tilde{g}_k$
```

For comparison with SIGNSGD theory, here we present non-convex convergence rates for SGD. These are classical results and we are not sure of the earliest reference.

We noticeably get exactly the same rate for large and small batch SGD when measuring convergence in terms of number of stochastic gradient calls. Although the rates are the same for a given number of gradient calls N, the large batch setting is preferred (in theory) since it achieves these N gradient calls in only $\sqrt{N}$ iterations, whereas the small batch setting requires N iterations. Fewer iterations in the large batch case implies a smaller wall-clock time to reach a given accuracy (assuming the large batch can be parallelised) as well as fewer rounds of communication in the distributed setting. These systems benefits of large batch learning have been observed by practitioners (Goyal et al., 2017).

Theorem C.1 (Non-convex convergence rate of SGD)**.** *Run algorithm C.1 for K iterations under Assumptions 1 to 3. Define $L := \|L\|_\infty$ and $\sigma^2 := \|\vec{\sigma}\|_2^2$. Set the learning rate and mini-batch size (independently of step k) as either*

$$\textbf{\textit{(large batch)}} \qquad \delta_k = \frac{1}{L} \qquad\qquad n_k = K$$

$$\textbf{\textit{(small batch)}} \qquad \delta_k = \frac{1}{L\sqrt{K}} \qquad\qquad n_k = 1$$

Let N be the cumulative number of stochastic gradient calls up to step K, i.e. $N = O(K^2)$ for large batch and $N = O(K)$ for small batch. Then, in either case, we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \leq \frac{1}{\sqrt{N}}\left[2L(f_0 - f_*) + \sigma^2\right]$$

Proof. The proof begins the same for the large and small batch case.

First we bound the improvement of the objective during a single step of the algorithm for one instantiation of the noise. g_k denotes the true gradient at step k and $\tilde{g}_k$ is a stochastic sample obeying Assumption 3.

Take Assumption 2 and plug in the algorithmic step.

$$\begin{aligned}
f_{k+1} - f_k &\leq g_k^T(x_{k+1} - x_k) + \sum_{i=1}^{d}\frac{L_i}{2}(x_{k+1} - x_k)_i^2 \\
&\leq g_k^T(x_{k+1} - x_k) + \frac{\|L\|_\infty}{2}\|x_{k+1} - x_k\|_2^2 \\
&= -\delta_k g_k^T\tilde{g}_k + \delta_k^2\frac{L}{2}\|\tilde{g}_k\|_2^2
\end{aligned}$$

Next we find the expected improvement at time $k+1$ conditioned on the previous iterate.

$$\mathbb{E}[f_{k+1} - f_k | x_k] \leq -\delta_k\|g_k\|_2^2 + \delta_k^2\frac{L}{2}\left(\sigma_k^2 + \|g_k\|_2^2\right).$$

σ_k^2 refers to the variance of the k^{th} stochastic gradient estimate, computed over a mini-batch of size n_k. Therefore, by Assumption 3, we have that $\sigma_k^2 \leq \sigma^2/n_k$.

First let's substitute in the **(large batch)** hyperparameters.

$$\begin{aligned}
\mathbb{E}[f_{k+1} - f_k | x_k] &\leq -\frac{1}{L}\|g_k\|_2^2 + \frac{1}{2L}\left(\frac{\sigma^2}{K} + \|g_k\|_2^2\right) \\
&= -\frac{1}{2L}\|g_k\|_2^2 + \frac{1}{2L}\frac{\sigma^2}{K}.
\end{aligned}$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the iterations:

$$\begin{aligned}
f_0 - f^* &\geq f_0 - \mathbb{E}[f_K] \\
&= \mathbb{E}\left[\sum_{k=0}^{K-1} f_k - f_{k+1}\right] \\
&\geq \frac{1}{2L}\mathbb{E}\sum_{k=0}^{K-1}\left[\|g_k\|_2^2 - \sigma^2\right]
\end{aligned}$$

We can rearrange this inequality to yield the rate:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \leq \frac{1}{K}\left[2L(f_0 - f_*) + \sigma^2\right].$$

Since we are growing our mini-batch size, it will take $N = O(K^2)$ gradient calls to reach step K. Substitute this in and we are done for the **(large batch)** case.

Now we need to show that the same result holds for the **(small batch)** case. Following the initial steps of the large batch proof, we get

$$\mathbb{E}[f_{k+1} - f_k | x_k] \leq -\delta_k \|g_k\|_2^2 + \delta_k^2 \frac{L}{2}\left(\sigma_k^2 + \|g_k\|_2^2\right).$$

This time $\sigma_k^2 = \sigma^2$. Substituting this and our learning rate and mini-batch settings into the expected improvement:

$$\begin{aligned}\mathbb{E}[f_{k+1} - f_k | x_k] &\leq -\frac{1}{L\sqrt{K}}\|g_k\|_2^2 + \frac{1}{2LK}\left(\sigma^2 + \|g_k\|_2^2\right)\\ &\leq -\frac{1}{2L\sqrt{K}}\|g_k\|_2^2 + \frac{1}{2L}\frac{\sigma^2}{K}.\end{aligned}$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the iterations:

$$\begin{aligned}f_0 - f^* &\geq f_0 - \mathbb{E}[f_K]\\ &= \mathbb{E}\left[\sum_{k=0}^{K-1} f_k - f_{k+1}\right]\\ &\geq \frac{1}{2L}\mathbb{E}\sum_{k=0}^{K-1}\left[\frac{\|g_k\|_2^2}{\sqrt{K}} - \frac{\sigma^2}{K}\right].\end{aligned}$$

We can rearrange this inequality to yield the rate:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \leq \frac{1}{\sqrt{K}}\left[2L(f_0 - f_*) + \sigma^2\right]$$

It will take $N = O(K)$ gradient calls to reach step K. Substitute this in and we are done. □

D. Proving the convergence rate of distributed SIGNSGD with majority vote

Theorem 2 (Non-convex convergence rate of distributed SIGNSGD with majority vote)**.** *Run algorithm 3 for K iterations under Assumptions 1 to 3. Set the learning rate and mini-batch size for each worker (independently of step k) as*

$$\delta_k = \frac{1}{\sqrt{\|\vec{L}\|_1 K}} \qquad\qquad n_k = K$$

*Then **(a)** majority vote with M workers converges at least as fast as* SIGNSGD *in Theorem 1.*

*And **(b)** further assuming that the noise in each component of the stochastic gradient is unimodal and symmetric about the mean (e.g. Gaussian), majority vote converges at improved rate:*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_1\right]^2\\ &\qquad\leq \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + \frac{2}{\sqrt{M}}\|\vec{\sigma}\|_1\right]^2\end{aligned}$$

where N is the cumulative number of stochastic gradient calls per worker up to step K.

Before we introduce the unimodal symmetric assumption, let's first address the claim that M-worker majority vote is at least as good as single-worker SIGNSGD as in Theorem 1 only using Assumptions 1 to 3.

Proof of (a). Recall that a crucial step in Theorem 1 is showing that

$$|g_i|\, \mathbb{P}[\text{sign}(\tilde{g}_i) \neq \text{sign}(g_i)] \leq \sigma_i$$

for component i of the stochastic gradient with variance bound σ_i.

The only difference in majority vote is that instead of using $\text{sign}(\tilde{g}_i)$ to approximate $\text{sign}(g_i)$, we are instead using $\text{sign}\left[\sum_{m=1}^{M} \text{sign}(\tilde{g}_{m,i})\right]$. If we can show that the same bound in terms of σ_i holds instead for

$$|g_i|\, \mathbb{P}\left[\text{sign}\left[\sum_{m=1}^{M} \text{sign}(\tilde{g}_{m,i})\right] \neq \text{sign}(g_i)\right] \tag{$\star$}$$

then we are done, since the machinery of Theorem 1 can then be directly applied.

Define the signal-to-noise ratio of a component of the stochastic gradient as $S := |g_i|/\sigma_i$. Note that when $S \leq 1$ then ($\star$) is trivially satisfied, so we need only consider the case that $S > 1$. S should really be labeled S_i but we abuse notation.

Without loss of generality, assume that g_i is negative, and thus using Assumption 3 and Cantelli's inequality (Cantelli, 1928) we get that for the failure probability q of a single worker

$$q := \mathbb{P}[\text{sign}(\tilde{g}_i) \neq \text{sign}(g_i)] = \mathbb{P}[\tilde{g}_i - g_i \geq |g_i|] \leq \frac{1}{1 + \frac{g_i^2}{\sigma_i^2}}$$

For $S > 1$ then we have failure probability $q < \frac{1}{2}$. If the failure probability of a single worker is smaller than $\frac{1}{2}$ then the server is essentially receiving a repetition code R_M of the true gradient sign. Majority vote is the maximum likelihood decoder of the repetition code, and of course decreases the probability of error—see e.g. (MacKay, 2002). Therefore in all regimes of S we have that

$$(\star) \leq |g_i|\, \mathbb{P}[\text{sign}(\tilde{g}_i) \neq \text{sign}(g_i)] \leq \sigma_i$$

and we are done. □

That's all well and good, but what we'd really like to show is that using M workers provides a speedup by reducing the variance. Is

$$(\star) \overset{?}{\leq} \frac{\sigma_i}{\sqrt{M}} \tag{$\dagger$}$$

too much to hope for?

Well in the regime where $S \gg 1$ such a speedup is very reasonable since $q \ll \frac{1}{2}$ by Cantelli, and the repetition code actually supplies exponential reduction in failure rate. But we need to exclude very skewed or bimodal distributions where $q > \frac{1}{2}$ and adding more voting workers will not help. That brings us naturally to the following lemma:

Lemma D.1 (Failure probability of a sign bit under conditions of unimodal symmetric gradient noise)**.** *Let $\tilde{g}_i$ be an unbiased stochastic approximation to gradient component g_i, with variance bounded by σ_i^2. Further assume that the noise distribution is unimodal and symmetric. Define signal-to-noise ratio $S := \frac{|g_i|}{\sigma_i}$. Then we have that*

$$\mathbb{P}[\textit{sign}(\tilde{g}_i) \neq \textit{sign}(g_i)] \leq \begin{cases} \frac{2}{9}\frac{1}{S^2} & \textit{if } S > \frac{2}{\sqrt{3}}, \\ \frac{1}{2} - \frac{S}{2\sqrt{3}} & \textit{otherwise} \end{cases}$$

which is in all cases less than $\frac{1}{2}$.

Proof. Recall Gauss' inequality for unimodal random variable X with mode ν and expected squared deviation from the mode τ^2 (Gauss, 1823; Pukelsheim, 1994):

$$\mathbb{P}[|X-\nu|>k] \leq \begin{cases} \frac{4}{9}\frac{\tau^2}{k^2} & \text{if } \frac{k}{\tau} > \frac{2}{\sqrt{3}}, \\ 1-\frac{k}{\sqrt{3}\tau} & \text{otherwise} \end{cases}$$

By the symmetry assumption, the mode is equal to the mean, so we replace mean $\mu = \nu$ and variance $\sigma^2 = \tau^2$.

$$\mathbb{P}[|X-\mu|>k] \leq \begin{cases} \frac{4}{9}\frac{\sigma^2}{k^2} & \text{if } \frac{k}{\sigma} > \frac{2}{\sqrt{3}}, \\ 1-\frac{k}{\sqrt{3}\sigma} & \text{otherwise} \end{cases}$$

Without loss of generality assume that g_i is negative. Then applying symmetry followed by Gauss, the failure probability for the sign bit satisfies:

$$\begin{aligned} \mathbb{P}[\text{sign}(\tilde{g}_i) \neq \text{sign}(g_i)] &= \mathbb{P}[\tilde{g}_i - g_i \geq |g_i|] \\ &= \frac{1}{2}\mathbb{P}[|\tilde{g}_i - g_i| \geq |g_i|] \\ &\leq \begin{cases} \frac{2}{9}\frac{\sigma_i^2}{g_i^2} & \text{if } \frac{|g_i|}{\sigma} > \frac{2}{\sqrt{3}}, \\ \frac{1}{2} - \frac{|g_i|}{2\sqrt{3}\sigma_i} & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{2}{9}\frac{1}{S^2} & \text{if } S > \frac{2}{\sqrt{3}}, \\ \frac{1}{2} - \frac{S}{2\sqrt{3}} & \text{otherwise} \end{cases} \end{aligned}$$

□

We now have everything we need to prove part (b) of Theorem 2.

Proof of (b). If we can show (†) we'll be done, since the machinery of Theorem 1 follows through with σ replaced everywhere by $\frac{\sigma}{\sqrt{M}}$. Note that the important quantity appearing in $(\star)$ is

$$\sum_{m=1}^{M} \text{sign}(\tilde{g}_{m,i}).$$

Let Z count the number of workers with correct sign bit. To ensure that

$$\text{sign}\left[\sum_{m=1}^{M} \text{sign}(\tilde{g}_{m,i})\right] = \text{sign}(g_i)$$

Z must be larger than $\frac{M}{2}$. But Z is the sum of M independent Bernoulli trials, and is therefore binomial with success probability p and failure probability q to be determined. Therefore we have reduced proving (†) to showing that

$$\mathbb{P}\left[Z \leq \frac{M}{2}\right] \leq \frac{1}{\sqrt{M}S} \tag{§}$$

where Z is the number of successes of a binomial random variable $b(M,p)$ and S is our signal-to-noise ratio $S := \frac{|g_i|}{\sigma_i}$.

Let's start by getting a bound on the success probability p (or equivalently failure probability q) of a single Bernoulli trial.

By Lemma D.1, which critically relies on unimodal symmetric gradient noise, the failure probability for the sign bit of a single worker satisfies:

$$\begin{aligned} q &:= \mathbb{P}[\text{sign}(\tilde{g}_i) \neq \text{sign}(g_i)] \\ &\leq \begin{cases} \frac{2}{9}\frac{1}{S^2} & \text{if } S > \frac{2}{\sqrt{3}}, \\ \frac{1}{2} - \frac{S}{2\sqrt{3}} & \text{otherwise} \end{cases} \\ &:= \tilde{q}(S) \end{aligned}$$

Where we have defined $\tilde{q}(S)$ to be our S-dependent bound on q. Since $q \leq \tilde{q}(S) < \frac{1}{2}$, there is hope to show $(\dagger)$. Define ϵ to be the defect of q from one half, and let $\tilde{\epsilon}(S)$ be its S-dependent bound.

$$\epsilon := \frac{1}{2} - q = p - \frac{1}{2} \geq \frac{1}{2} - \tilde{q}(S) := \tilde{\epsilon}(S)$$

Now we have an analytical handle on random variable Z, we may proceed to show $(\S)$. There are a number of different inequalities that we can use to bound the tail of a binomial random variable, but Cantelli's inequality will be good enough for our purposes.

Let $\bar{Z} := M - Z$ denote the number of failures. $\bar{Z}$ is binomial with mean $\mu_{\bar{Z}} = Mq$ and variance $\sigma^2_{\bar{Z}} = Mpq$. Then using Cantelli we get

$$\begin{aligned}
\mathbb{P}\left[Z \leq \frac{M}{2}\right] &= \mathbb{P}\left[\bar{Z} \geq \frac{M}{2}\right] \\
&= \mathbb{P}\left[\bar{Z} - \mu_{\bar{Z}} \geq \frac{M}{2} - \mu_{\bar{Z}}\right] \\
&= \mathbb{P}\left[\bar{Z} - \mu_{\bar{Z}} \geq M\epsilon\right] \\
&\leq \frac{1}{1 + \frac{M^2\epsilon^2}{Mpq}} \\
&\leq \frac{1}{1 + \frac{M}{\frac{1}{4\epsilon^2} - 1}}
\end{aligned}$$

Now using the fact that $\frac{1}{1+x^2} \leq \frac{1}{2x}$ we get

$$\mathbb{P}\left[Z \leq \frac{M}{2}\right] \leq \frac{\sqrt{\frac{1}{4\epsilon^2} - 1}}{2\sqrt{M}}$$

To finish, we need only show that $\sqrt{\frac{1}{4\epsilon^2} - 1}$ is smaller than $\frac{2}{S}$, or equivalently that its square is smaller than $\frac{4}{S^2}$. Well plugging in our bound on ϵ we get that

$$\frac{1}{4\epsilon^2} - 1 \leq \frac{1}{4\tilde{\epsilon}(S)^2} - 1$$

where

$$\tilde{\epsilon}(S) = \begin{cases} \frac{1}{2} - \frac{2}{9}\frac{1}{S^2} & \text{if } S > \frac{2}{\sqrt{3}}, \\ \frac{S}{2\sqrt{3}} & \text{otherwise} \end{cases}$$

First take the case $S \leq \frac{2}{\sqrt{3}}$. Then $\tilde{\epsilon}^2 = \frac{S^2}{12}$ and $\frac{1}{4\tilde{\epsilon}^2} - 1 = \frac{3}{S^2} - 1 < \frac{4}{S^2}$. Now take the case $S > \frac{2}{\sqrt{3}}$. Then $\tilde{\epsilon} = \frac{1}{2} - \frac{2}{9}\frac{1}{S^2}$ and we have $\frac{1}{4\tilde{\epsilon}^2} - 1 = \frac{1}{S^2}\frac{\frac{8}{9} - \frac{16}{81}\frac{1}{S^2}}{1 - \frac{8}{9}\frac{1}{S^2} + \frac{16}{81}\frac{1}{S^4}} < \frac{1}{S^2}\frac{\frac{8}{9}}{1 - \frac{8}{9}\frac{1}{S^2}} < \frac{4}{S^2}$ by the condition on S.

So we have shown both cases, which proves $(\S)$ from which we get $(\dagger)$ and we are done. □

E. General recipes for the convergence of approximate sign gradient methods

Now we generalize the arguments in the proof of SIGNSGD and prove a master lemma that provides a general recipe for analyzing the approximate sign gradient method. This allows us to handle momentum and the majority voting schemes, hence proving Theorem 3 and Theorem 2.

Lemma E.1 (Convergence rate for a class of approximate sign gradient method)**.** *Let C and K be integers satisfying $0 < C \ll K$. Consider the algorithm given by $x_{k+1} = x_k - \delta_k sign(v_k)$, for a fixed positive sequence of δ_k and where $v_k \in \mathbb{R}^d$ is a measurable and square integrable function of the entire history up to time k, including $x_1, ..., x_k, v_1, ..., v_{k-1}$*

and all N_k stochastic gradient oracle calls up to time k. Let $g_k = \nabla f(x_k)$. If Assumption 1 and Assumption 2 are true and in addition for $k = C, C+1, C+2, ..., K$

$$\mathbb{E}\left[\sum_{i=1}^{d} |g_{k,i}| \mathbb{P}[sign(v_{k,i}) \neq sign(g_{k,i}) | x_k]\right] \leq \xi(k) \tag{2}$$

where the expectation is taken over the all random variables, and the rate $\xi(k)$ obeys that $\xi(k) \to 0$ as $k \to \infty$ and then we have

$$\frac{1}{K-C}\sum_{k=C}^{K-1} \mathbb{E}\|g_k\|_1 \leq \frac{f_C - f_* + 2\sum_{k=C}^{K-1}\delta_k \xi(k) + \sum_{k=C}^{K-1}\frac{\delta_k^2\|\vec{L}\|_1}{2}}{(K-C)\min_{C\leq k\leq K-1}\delta_k}.$$

In particular, if $\delta_k = \delta/\sqrt{k}$ and $\xi(k) = \kappa/\sqrt{k}$, for some problem dependent constant κ, then we have

$$\frac{1}{K-C}\sum_{k=C}^{K-1} \mathbb{E}\|g_k\|_1 \leq \frac{\frac{f_C - f_*}{\delta} + (2\kappa + \|\vec{L}\|_1\delta/2)(\log K + 1)}{\sqrt{K} - \frac{C}{\sqrt{K}}}.$$

Proof. Our general strategy will be to show that the expected objective improvement at each step will be good enough to guarantee a convergence rate in expectation. First let's bound the improvement of the objective during a single step of the algorithm for $k \geq C$, and then take expectation. Note that $\mathbb{I}[.]$ is the indicator function, and $w_k[i]$ denotes the i^{th} component of the vector w_k.

By Assumption 2

$$\begin{aligned} f_{k+1} - f_k &\leq g_k^T(x_{k+1} - x_k) + \sum_{i=1}^{d}\frac{L_i}{2}(x_{k+1,i} - x_{k,i})^2 && \text{Assumption 2} \\ &= -\delta_k g_k^T \text{sign}(v_k) + \delta_k^2\frac{\|\vec{L}\|_1}{2} && \text{by the update rule} \\ &= -\delta_k\|g_k\|_1 + 2\delta_k\sum_{i=1}^{d}|g_k[i]|\,\mathbb{I}[\text{sign}(v_k[i]) \neq \text{sign}(g_k[i])] + \delta_k^2\frac{\|\vec{L}\|_1}{2} && \text{by identity} \end{aligned}$$

Now, for $k \geq C$ we need to find the expected improvement at time $k+1$ conditioned on x_k, where the expectation is over the randomness of the stochastic gradient oracle. Note that $\mathbb{P}[E]$ denotes the probability of event E.

$$\mathbb{E}[f_{k+1} - f_k | x_k] \leq -\delta_k\|g_k\|_1 + 2\delta_k\sum_{i=1}^{d}|g_k[i]|\,\mathbb{P}\Big[\text{sign}(v_k[i]) \neq \text{sign}(g_k[i])\Big| x_k\Big] + \delta_k^2\frac{\|\vec{L}\|_1}{2}.$$

Note that g_k becomes fixed when we condition on x_k. Further take expectation over x_k, and apply (2). We get:

$$\mathbb{E}[f_{k+1} - f_k] \leq -\delta_k\mathbb{E}[\|g_k\|_1] + 2\delta_k\xi(k) + \frac{\delta_k^2\|\vec{L}\|_1}{2}. \tag{3}$$

Rearrange the terms and sum over (3) for $k = C, C+1, ..., K-1$.

$$\sum_{k=C}^{K-1}\delta_k\mathbb{E}[\|g_k\|_1] \leq \sum_{k=C}^{K-1}(\mathbb{E}f_k - \mathbb{E}f_{k+1}) + 2\sum_{k=C}^{K-1}\delta_k\xi(k) + \sum_{k=C}^{K-1}\frac{\delta_k^2\|\vec{L}\|_1}{2}$$

Dividing both sides by $\left[(K-C)\min_{C\leq k\leq K-1}\delta_k\right]$, using a telescoping sum over $\mathbb{E}f_k$ and using that $f(x) \geq f_*$ for all x, we get

$$\frac{1}{K-C}\sum_{k=C}^{K-1} \mathbb{E}\|g_k\|_1 \leq \frac{f_C - f_* + 2\sum_{k=C}^{K-1}\delta_k \xi(k) + \sum_{k=C}^{K-1}\frac{\delta_k^2\|\vec{L}\|_1}{2}}{(K-C)\min_{C\leq k\leq K-1}\delta_k}$$

and the proof is complete by noting that the minimum is smaller than the average in the LHS. □

To use the above Lemma for analyzing SIGNUM and the Majority Voting scheme, it suffices to check condition (2) for each algorithm.

One possible way to establish (2) is show that v_k is a good approximation of the gradient g_k in expected absolute value.

Lemma E.2 (Estimation to testing reduction)**.** *Equation* (2) *is true, if for every* k

$$\sum_{i=1}^{d} \mathbb{E}|v_k[i] - g_k[i]| \leq \xi(k). \tag{4}$$

Proof. First note that for any two random variables $a, b \in \mathbb{R}$.

$$\mathbb{P}\Big[\text{sign}(a) \neq \text{sign}(b)\Big] \leq \mathbb{P}\Big[|a-b| > |b|\Big].$$

Condition on x_k and apply the above inequality to every $i = 1, ..., d$ to what is inside the expectation of (2), we have

$$\sum_{i=1}^{d} |g_k[i]|\mathbb{P}\Big[\text{sign}(v_k[i]) \neq \text{sign}(g_k[i])\Big|x_k\Big] \leq \sum_{i=1}^{d} |g_k[i]|\mathbb{P}\Big[|v_k[i] - g_k[i]| > |g_k[i]|\Big|x_k\Big] \leq \sum_{i=1}^{d} \mathbb{E}[|v_k[i] - g_k[i]||x_k].$$

Note that the final $\leq$ uses Markov's inequality and constant $|g_k[i]|$ cancels out.

The proof is complete by taking expectation on both sides and apply (4). □

Note that the proof of this lemma uses Markov's inequality in the same way information-theoretical lower bounds are often proved in statistics — reducing estimation to testing.

Another handy feature of the result is that we do not require the approximation to hold for every possible x_k. It is okay that for some x_k, the approximation is much worse as long as those x_k appears with small probability according to the algorithm. This feature enables us to analyze momentum and hence proving the convergence for SIGNUM.

F. Analysis for SIGNUM

Recall our definition of the key random variables used in SIGNUM.

$$\begin{aligned} g_k &:= \nabla f(x_k) \\ \tilde{g}_k &:= \frac{1}{n_k}\sum_{j=1}^{n_k} \tilde{g}^{(j)}(x_k) \\ m_k &:= \frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\big[\beta^t g_{k-t}\big] \\ \tilde{m}_k &:= \frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\big[\beta^t \tilde{g}_{k-t}\big] \end{aligned}$$

SIGNUM effectively uses $v_k = \tilde{m}_k$ and also $\delta_k = O(1/\sqrt{k})$.

Before we prove the convergence of SIGNUM, we first prove a utility lemma about the random variable $Z_k := \tilde{g}_k - g_k$. Note that in this lemma quantities like Z_k, Y_k, $|Z_k|$ and Z_k^2 are considered vectors—so this lemma is a statement about each component of the vectors separately and all operations, such as $(\cdot)^2$ are pointwise operations.

Lemma F.1 (Cumulative error of stochastic gradient)**.** *For any* $k < \infty$ *and fixed weight* $-\infty < \alpha_1, ..., \alpha_k < \infty$, $\sum_{l=1}^{k} \alpha_l Z_l$ *is a Martingale. In particular,*

$$\mathbb{E}\left[\left(\sum_{l=1}^{k} \alpha_l Z_l\right)^2\right] \leq \sum_{l=1}^{k} \alpha_l^2 \vec{\sigma}^2.$$

Proof. We simply check the definition of a Martingale. Denote $Y_k := \sum_{l=1}^k \alpha_l Z_l$. First, we have that

$$\begin{aligned}
\mathbb{E}[|Y_k|] &= \mathbb{E}\left[\left|\sum_{l=1}^k \alpha_l Z_l\right|\right] \\
&\leq \sum_l |\alpha_l| \mathbb{E}[|Z_l|] && \text{triangle inequality} \\
&= \sum_l |\alpha_l| \mathbb{E}\Big[\mathbb{E}[|Z_l||x_l]\Big] && \text{law of total probability} \\
&\leq \sum_l |\alpha_l| \mathbb{E}\left[\sqrt{\mathbb{E}[Z_l^2|x_l]}\right] && \text{Jensen's inequality} \\
&\leq \sum_l |\alpha_l| \vec{\sigma} < \infty
\end{aligned}$$

Second, again using the law of total probability,

$$\begin{aligned}
\mathbb{E}[Y_{k+1}|Y_1, ..., Y_k] &= \mathbb{E}\left[\sum_{l=1}^{k+1} \alpha_l Z_l \middle| \alpha_1 Z_1, ..., \alpha_k Z_k\right] \\
&= Y_k + \alpha_{k+1} \mathbb{E}\left[Z_{k+1} | \alpha_1 Z_1, ..., \alpha_k Z_k\right] \\
&= Y_k + \alpha_{k+1} \mathbb{E}\left[\mathbb{E}\left[Z_{k+1} | x_{k+1}, \alpha_1 Z_1, ..., \alpha_k Z_k\right] | \alpha_1 Z_1, ..., \alpha_k Z_k\right] \\
&= Y_k + \alpha_{k+1} \mathbb{E}\left[\mathbb{E}\left[Z_{k+1} | x_{k+1}\right] | \alpha_1 Z_1, ..., \alpha_k Z_k\right] \\
&= Y_k
\end{aligned}$$

This completes the proof that it is indeed a Martingale. We now make use of the properties of Martingale difference sequences to establish a variance bound on the Martingale.

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{l=1}^k \alpha_l Z_l\right)^2\right] &= \sum_{l=1}^k \mathbb{E}[\alpha_l^2 Z_l^2] + 2\sum_{l<j} \mathbb{E}[\alpha_l \alpha_j Z_l Z_j] \\
&= \sum_{l=1}^k \alpha_l^2 \mathbb{E}[\mathbb{E}[Z_l^2 | Z_1, ..., Z_{l-1}]] + 2\sum_{l<j} \alpha_l \alpha_j \mathbb{E}\Big[Z_l \mathbb{E}\big[\mathbb{E}[Z_j | Z_1, ..., Z_{j-1}]\big| Z_l\big]\Big] \\
&= \sum_{l=1}^k \alpha_l^2 \mathbb{E}[\mathbb{E}[\mathbb{E}[Z_l^2 | x_l, Z_1, ..., Z_{l-1}] | Z_1, ..., Z_{l-1}]] + 0 \\
&= \sum_{l=1}^k \alpha_l^2 \vec{\sigma}^2.
\end{aligned}$$

□

The consequence of this lemma is that we are able to treat $Z_1, ..., Z_k$ as if they are independent, even though they are not—clearly Z_l is dependent on $Z_1, ..., Z_{l-1}$ through x_l.

Lemma F.2 (Gradient approximation in SIGNUM)**.** *The version of the* SIGNUM *algorithm that takes $v_k = \tilde{m}_k$, and all parameters according to Theorem 3, obeys that for all integer $C \leq k \leq K$*

$$||\mathbb{E}|\tilde{m}_k - g_k|||_1 \leq \frac{2}{\sqrt{k+1}} \left(8\|\vec{L}\|_1 \delta \frac{\beta}{1-\beta} + \sqrt{3}\|\vec{\sigma}\|_1 \sqrt{1-\beta}\right).$$

Proof. For each $i \in [d]$ we will use the following non-standard "bias-variance" decomposition.

$$\mathbb{E}\left[|\tilde{m}_k[i] - g_k[i]|\right] \leq \underbrace{\mathbb{E}\left[|m_k[i] - g_k[i]|\right]}_{(*)} + \underbrace{\mathbb{E}\left[|\tilde{m}_k[i] - m_k[i]|\right]}_{(**)} \tag{5}$$

We will first bound $(**)$ and then deal with $(*)$.

Note that $(**) = \frac{1-\beta}{1-\beta^{k+1}}\mathbb{E}\left[|\sum_{t=0}^{k}\beta^{k-t}Z_t|\right]$. Using Jensen's inequality and applying Lemma F.1 with our choice of $\alpha_1, ..., \alpha_l$ (including the effect of the increasing batch size) we get that for $k \geq C$

$$(**) \leq \frac{1-\beta}{1-\beta^{k+1}}\sqrt{\mathbb{E}\left[\left|\sum_{t=0}^{k}\beta^t Z_{k-t}\right|^2\right]} \leq \frac{1-\beta}{1-\beta^{k+1}}\sqrt{\sum_{t=0}^{k}\left[\beta^{2t}\frac{\vec{\sigma}^2}{k-t+1}\right]},$$

where

$$\begin{aligned}
\sum_{t=0}^{k}\left[\beta^{2t}\frac{\vec{\sigma}^2}{k-t+1}\right] &= \sum_{t=0}^{\frac{k}{2}}\left[\beta^{2t}\frac{\vec{\sigma}^2}{k-t+1}\right] + \sum_{t=\frac{k}{2}+1}^{k}\left[\beta^{2t}\frac{\vec{\sigma}^2}{k-t+1}\right] && \text{break up sum}\\
&\leq \sum_{t=0}^{\frac{k}{2}}\left[\beta^{2t}\right]\frac{\vec{\sigma}^2}{\frac{k}{2}+1} + \sum_{t=\frac{k}{2}+1}^{k}\left[\beta^k\vec{\sigma}^2\right] && \text{bound summands}\\
&\leq \frac{1}{1-\beta^2}\frac{\vec{\sigma}^2}{\frac{k}{2}+1} + \frac{k}{2}\beta^k\vec{\sigma}^2 && \text{geometric series}\\
&\leq \frac{3}{1-\beta^2}\frac{\vec{\sigma}^2}{k+1} && \text{since } k \geq C
\end{aligned}$$

Combining, and again using our condition that $k \geq C$, we get

$$(**) \leq \frac{1-\beta}{1-\beta^{k+1}}\sqrt{\frac{3}{1-\beta^2}}\frac{\vec{\sigma}}{\sqrt{k+1}} \leq 2\sqrt{3}\sqrt{1-\beta}\frac{\vec{\sigma}}{\sqrt{k+1}} \tag{6}$$

We now turn to bounding $(*)$ — the "bias" term.

$$\begin{aligned}
\mathbb{E}[|m_k - g_k|] &= \mathbb{E}\left[\left|\frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\left[\beta^t g_{k-t}\right] - \frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\left[\beta^t g_k\right]\right|\right] && \text{since } \frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\beta^t = 1\\
&\leq \frac{1-\beta}{1-\beta^{k+1}}\sum_{t=0}^{k}\left[\beta^t\mathbb{E}[|g_{k-t}-g_k|]\right]\\
&\leq 2(1-\beta)\sum_{t=1}^{k}\beta^t\mathbb{E}[|g_{k-t}-g_k|] && \text{since } k \geq C
\end{aligned} \tag{7}$$

To proceed, we need the following lemma.

Lemma F.3. *Under Assumption 2, for any sign vector $s \in \{-1,1\}^d$, any $x \in \mathbb{R}^d$ and any $\epsilon \leq \delta$*

$$\|g(x+\epsilon s) - g(x)\|_1 \leq 2\epsilon\|\vec{L}\|_1.$$

Proof. By Taylor's theorem,

$$g(x+\epsilon s) - g(x) = \left[\int_{t=0}^{1} H(x+t\epsilon s)dt\right]\epsilon s.$$

Let $v := \text{sign}(g(x+\epsilon s) - g(x))$, $H := \left[\int_{t=0}^{1} H(x+t\epsilon s)dt\right]$ and moreover, use H_+ to denote the psd part of H and H_- to denote the nsd part of H. Namely, $H = H_+ - H_-$.

We can write

$$\begin{aligned}
\|g(x+\epsilon s) - g(x)\|_1 &= v^T(g(x+\epsilon s) - g(x)) = v^T H(\epsilon s) = \epsilon v^T H_+ s - \epsilon v^T H_- s\\
&= \epsilon\langle H_+^{1/2}v, H_+^{1/2}s\rangle - \epsilon\langle H_-^{1/2}v, H_-^{1/2}s\rangle \leq \epsilon\|H_+^{1/2}v\|\|H_+^{1/2}s\| + \epsilon\|H_-^{1/2}v\|\|H_-^{1/2}s\|.
\end{aligned} \tag{8}$$

Note that assumption 2 implies the semidefinite ordering

$$H_+ \prec \operatorname{diag}(\vec{L}) \text{ and } H_- \prec \operatorname{diag}(\vec{L})$$

and thus $\max\{s^T H_+ s, s^T H_- s\} \leq \sum_{i=1}^d L_i = \|\vec{L}\|_1$ for all $s \in \{-1,1\}^d$.

The proof is complete by observing that both v and s are sign vectors in (8). □

Using the above lemma and the fact that our update rules are always following some sign vectors with learning rate smaller than δ, we have

$$\begin{aligned}
\|g_{k-t} - g_k\|_1 &\leq \sum_{l=0}^{t-1} \|g_{k-l} - g_{k-l-1}\|_1 \\
&\leq 2\|\vec{L}\|_1 \sum_{l=0}^{t-1} \delta_{k-l-1} \leq \sum_{l=0}^{t-1} \frac{2\|\vec{L}\|_1 \delta}{\sqrt{k-l}} \\
&\leq 2\|\vec{L}\|_1 \delta \int_{k-t}^{k} \frac{dx}{\sqrt{x}} = 4\|\vec{L}\|_1 \delta \left(\sqrt{k} - \sqrt{k-t}\right) \\
&\leq 4\|\vec{L}\|_1 \delta \sqrt{k} \left(1 - \sqrt{1 - \frac{t}{k}}\right) \\
&\leq 4\|\vec{L}\|_1 \delta \frac{t}{\sqrt{k}} && \text{for } x \geq 0, 1 - x \leq \sqrt{1-x} \qquad (9)
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_i \mathbb{E}[|m_k[i] - g_k[i]|] = \mathbb{E}\left[\|m_k - g_k\|_1\right] \\
&\leq 2(1-\beta) \sum_{t=1}^{k} \beta^t \mathbb{E}\|g_{k-t} - g_k\|_1 && \text{Apply (7)} \\
&\leq \frac{8(1-\beta)\|\vec{L}\|_1 \delta}{\sqrt{k}} \sum_{t=1}^{\infty} t\beta^t && \text{Apply (9) and extend sum to } \infty \\
&\leq \frac{8(1-\beta)\|\vec{L}\|_1 \delta}{\sqrt{k}} \frac{\beta}{(1-\beta)^2} && \text{derivative of geometric progression} \\
&\leq \frac{16\|\vec{L}\|_1 \delta}{\sqrt{k+1}} \frac{\beta}{1-\beta} && \text{for } k \geq 1, \sqrt{\frac{k+1}{k}} \leq 2 \qquad (10)
\end{aligned}$$

Substitute (6) into (5), sum both sides over i and then further plug in (10) we get the statement in the lemma. □

The proof of Theorem 3 now follows in a straightforward manner. Note that Lemma F.2 only kicks in after a warmup period of C iterations, with C as specified in Theorem 3. In theory it does not matter what you do during this warmup period, provided you accumulate the momentum as normal and take steps according to the prescribed learning rate and mini-batch schedules. One option is to just stay put and not update the parameter vector for the first C iterations. This is wasteful since no progress will be made on the objective. A better option in practice is to take steps using the sign of the stochastic gradient (i.e. do SIGNSGD) instead of SIGNUM during the warmup period.

Proof of Theorem 3. Substitute Lemma F.2 as $\xi(k)$ into Lemma E.1 and check that $\xi(k) = O(1/\sqrt{k}), \delta_k = O(1/\sqrt{k}), \min \delta_k = \delta/\sqrt{K}, C \ll K$, and in addition, we note that by the increasing minibatch size $N_K = O(K^2)$. Substitute $K = O(\sqrt{N})$ and take the square on both sides of the inequality. (We can take the $\min_k$ out of the square since all the arguments are nonnegative and $(\cdot)^2$ is monotonic on $\mathbb{R}_+$). □