On the Interaction of Noise, Compression, and Adaptivity under (L_0, L_1)-Smoothness: An SDE Approach

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Using stochastic differential equation (SDE) approximations, we study the dynamics of Distributed SGD, Distributed Compressed SGD, and Distributed SignSGD under (L_0, L_1)-smoothness and flexible noise assumptions. Our analysis provides insights – which we validate through simulation – into the intricate interactions between batch noise, stochastic gradient compression, and adaptivity in this modern theoretical setup. For instance, we show that *adaptive* methods such as Distributed SignSGD can successfully converge under standard assumptions on the learning rate scheduler, even under heavy-tailed noise. On the contrary, Distributed (Compressed) SGD with pre-scheduled decaying learning rate fails to achieve convergence, unless such a schedule also accounts for an inverse dependency on the gradient norm – de facto falling back into an adaptive method.

1. Introduction

Understanding the dynamics of stochastic optimization algorithms is crucial, especially in distributed machine learning settings where batch noise, compression, and adaptivity significantly impact convergence and generalization. Despite extensive studies in the literature, the interplay among these three aspects under the general condition of (L_0, L_1)-smoothness remains underexplored.

Contributions. Our key contributions include:

- Establishing convergence bounds for Distributed SGD (DSGD), Distributed Compressed SGD (DCSGD), and Distributed SignSGD (DSignSGD) under the (L_0, L_1)-smoothness condition;
- Showcasing how normalizing the update step of DCSGD naturally emerges as a condition for convergence, thus confirming the superiority of adaptive methods for ill-conditioned loss landscapes, especially for pathological batch noise or when unbiased compression is used;
- Highlighting that an *adaptive* method such as DSignSGD converges even under heavy-tailed noise with standard assumptions on the learning rate scheduler.

2. Related work

SDE Approximations and Applications. In [23], a rigorous theoretical framework was introduced to derive SDEs that faithfully model the stochastic behavior intrinsic to optimization algorithms widely employed in machine learning. Since then, such SDE-based formulations have found application across several domains, including *stochastic optimal control* for tuning stepsizes [23, 24] and batch sizes [47]. Notably, SDEs have been instrumental in analyzing *convergence bounds* and *stationary distributions* [5, 6, 8], *scaling laws* [7, 8, 16], *implicit regularization* effects [5, 38], and *implicit preconditioning* [27, 44].

Interplay of noise, compression, and adaptivity under (L_0, L_1)-smoothness Previous research has extensively studied the effect of batch noise, compression, and adaptivity on the convergence of optimizers. Batch noise significantly influences stochastic gradient algorithms, affecting their convergence speed and stability [8, 20, 37, 45]. Noise characteristics such as heavy-tailed distributions have been shown to profoundly impact the optimization trajectories, necessitating robust algorithmic strategies [13, 36]. Compression methods, including unbiased techniques such as sparsification and quantization [1, 29, 39] and biased approaches like SignSGD [2, 3], are critical for reducing communication overhead in distributed training. These compression techniques come with theoretical guarantees under various smoothness assumptions [1, 7, 12, 29]. Adaptive methods such as SignSGD normalize gradient elements to cope effectively with large or heavy-tailed gradient noise, thus demonstrating improved empirical robustness [7, 8, 19, 34].

However, most of the aforementioned works rely on restrictive assumptions such as L-smoothness, i.e., the L-Lipschitz continuity of the gradient. To relax this condition, Zhang et al. [45] introduces and empirically validates the (L_0, L_1)-*smoothness* assumption, which allows the norm of the Hessian to be bounded by an affine function of the gradient norm, thereby significantly expanding the class of admissible problems. Various (stochastic) first-order methods have been analyzed under (L_0, L_1)-smoothness, including Clip-SGD and its variants [14, 18, 31, 40, 45, 46], Normalized SGD and its variants [4, 15, 48], SignSGD [9], AdaGrad [10, 43], Adam [22, 42], and SGD [21]. In the context of compressed communication, Khirirat et al. [17] proposed and analyzed a momentum-based variant of normalized EF21-SGD [32] under the assumption of bounded noise variance.

To the best of our knowledge, no study has jointly considered all these aspects, namely, batch noise, communication compression, and adaptivity, under the (L_0, L_1)-smoothness condition. In particular, we consider flexible noise assumptions ranging from bounded to unbounded variance, and even encompassing heavy-tailed noise. Our work closes this gap by providing a comprehensive analysis of their interplay within a unified theoretical framework.

3. Preliminaries

Distributed Setup. Let us consider the problem of minimizing an objective function expressed as an average of N functions: $\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{N}\sum_{i=1}^N f_i(x)\right]$, where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is lower bounded and twice continuously differentiable, and represents the loss over the local data of the i-th agent. In our stochastic setup, each agent only has access to gradient estimates: let n_i be the number of datapoints accessible to agent i; at a given $x \in \mathbb{R}^d$, agent i estimates $\nabla f_i(x)$ using a batch of data $\gamma_i \subseteq \{1, \dots, n_i\}$, sampled uniformly with replacement and uncorrelated from the previously sampled batches. Given the sampling properties above, this estimate, which we denote by $\nabla f_{i,\gamma_i}(x)$, can be modeled as a perturbation of the global gradient: $\nabla f_{i,\gamma_i}(x) = \nabla f(x) + Z_i(x)$.

Noise assumptions. We assume the sampling process and agent configurations are such that, for all $x \in \mathbb{R}^d$ and each agent pair (i, j) with $i \neq j$, $Z_i(x)$ is independent of $Z_j(x)$. Regarding assumptions on the noise structure, we always assume that at each $x \in \mathbb{R}^d$, $Z_i(x)$ is absolutely continuous and with coordinate-wise symmetric distribution. If we discuss the setting $Z_i(x) \in L^1(\mathbb{R}^d)$, then we assume $\mathbb{E}[Z_i(x)] = 0$. Last, if $Z_i(x) \in L^2(\mathbb{R}^d)$, we denote $\Sigma_i(x) := Cov(Z_i(x))$.

Next, we define our two structural assumptions. The first one strictly concerns the global landscape; the second concerns how global landscape features affect the noise distribution of each agent.

Definition 1 ([45]) *f is (L_0, L_1)-smooth ($L_0, L_1 \geq 0$) if, $\forall x \in \mathbb{R}^d$, $\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \|\nabla f(x)\|$.*

Definition 2 (Mod. of the assumptions from [35, 41]) *The gradient noise for agent i has $(\sigma_{0,i}^2, \sigma_{1,i}^2)$-variance if $\|\Sigma_i(x)\|_\infty \leq \sigma_{0,i}^2 + \sigma_{1,i}^2 \|\nabla f(x)\|_2^2$. If $\sigma_{1,i} = 0$, the noise has bounded variance.*

SDE approximations. The following definition formalizes the idea that an SDE can be a "reliable surrogate" to model an optimizer. It is drawn from the field of numerical analysis of SDEs (see [28]) and it quantifies the disparity between the discrete and the continuous processes.

Definition 3 *A continuous-time stochastic process $(X_t)_{t\in[0,T]}$ is an order α weak approximation of a discrete stochastic process $(x_k)_{k=0}^{\lfloor T/\eta \rfloor}$ if for every polynomial growth function g, there exists a positive constant C, independent of η, such that $\max_{k=0,\dots,\lfloor T/\eta \rfloor} |\mathbb{E} g(x_k) - \mathbb{E} g(X_{k\eta})| \leq C\eta^\alpha$.*

Optimizers and SDEs. We study: 1) DSGD defined as $x_{k+1} = x_k - \frac{\eta}{N} \sum_{i=1}^N \nabla f_{i,\gamma_i}(x_k)$ and whose SDE is defined in Eq. 8 (see Thm. 3.2 in [7]); 2) DCSGD defined as $x_{k+1} = x_k - \frac{\eta}{N} \sum_{i=1}^N \mathcal{C}_{\xi_i}(\nabla f_{i,\gamma_i}(x_k))$, where the stochastic compressors $\mathcal{C}_{\xi_i}$ are independent for different i and satisfy (i) $\mathbb{E}_{\xi_i}[\mathcal{C}_{\xi_i}(x)] = x$ and (ii) $\mathbb{E}_{\xi_i}\left[\|\mathcal{C}_{\xi_i}(x) - x\|_2^2\right] \leq \omega_i \|x\|_2^2$ for some compression rates $\omega_i \geq 0$: Its SDE is defined in Eq. 27 (see Thm. 3.6 in [7]); 3) DSignSGD defined as $x_{k+1} = x_k - \frac{\eta}{N} \sum_{i=1}^N \operatorname{sign}(\nabla f_{i,\gamma_i}(x_k))$ and whose SDE is in Eq. 51 (see Thm. 3.10 in [7]).

Importantly, extensive experimental validation [6–8, 26, 30] shows that the SDEs do track their respective optimizers accurately on a variety of architectures, e.g., MLPs, ResNets, and ViTs.

4. Theoretical Results

Recall that, in the continuous-time setup, the dynamics of the iterates is modeled by a stochastic process X_t solution to an SDE model. In this setting, the learning rate is a scalar factor in the SDE influencing both its drift and its diffusion. To decouple adaptivity from scheduling, we *parametrize our learning rate as a product*: $\eta\eta_t$. To ensure convergence, we **always** assume η_t satisfying the Robbins and Monro [33] conditions: For $\phi_t^i = \int_0^t (\eta_s)^i ds$, we require $\phi_t^1 \overset{t\to\infty}{\to} \infty$, $\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0$.

4.1. Overview

Under (L_0, L_1)-smoothness, our insights concern the structure of η for convergence, where $\eta\eta_t$ is the actual learning rate and η_t is a predetermined scheduler: See Fig. 1 for empirical validation.

- Thm. 4 shows that, if at least one agent i has $(\sigma_{0,i}^2, \sigma_{1,i}^2)$-variance, the dynamics of the DSGD model can converge to a first order stationary point in expectation, yet the learning rate η_t is required to also scale inversely proportional to the gradient – i.e. needs to be adaptive;
- Thm. 5 operates in the compressed unbiased gradient setting. The insights are similar to Thm. 4 yet assume bounded variance for pedagogical purposes only: Thm. 6 covers the more general $(\sigma_{0,i}^2, \sigma_{1,i}^2)$-variance case;
- Thm. 7 shows that the DSignSGD model does not require adaptive η to converge: Not even when the expectation of the batch noise is **unbounded** – The intuition is that DSignSGD is already normalized.

4.2. Results

We state the SDE models directly in the appendix and indicate the setting with blue color.

Theorem 4 ***(DSGD, unbounded variance)*** *Let f be (L_0, L_1)-smooth, and each agent have $(\sigma_{0,i}^2, \sigma_{1,i}^2)$-variance. Define $\overline{\sigma_0^2} := \frac{1}{N} \sum_{i=1}^N \sigma_{0,i}^2$ and $\overline{\sigma_1^2} := \frac{1}{N} \sum_{i=1}^N \sigma_{1,i}^2$. For an arbitrary $\epsilon \in (0,1)$, assume*

$$\eta\eta_t \leq \frac{2N\epsilon}{d\left(\overline{\sigma_1^2} L_0 + \overline{\sigma_0^2} L_1 + L_1 \overline{\sigma_1^2} \mathbb{E}[\|\nabla f(X_t)\|_2)\right]}. \tag{1}$$

Then, for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1 (1-\epsilon)} \left(f(X_0) - f(X_*) + \phi_t^2 \frac{\eta d (L_0 + L_1)(\overline{\sigma_0^2} + \overline{\sigma_1^2})}{2N} \right) \overset{t\to\infty}{\to} 0. \tag{2}$$

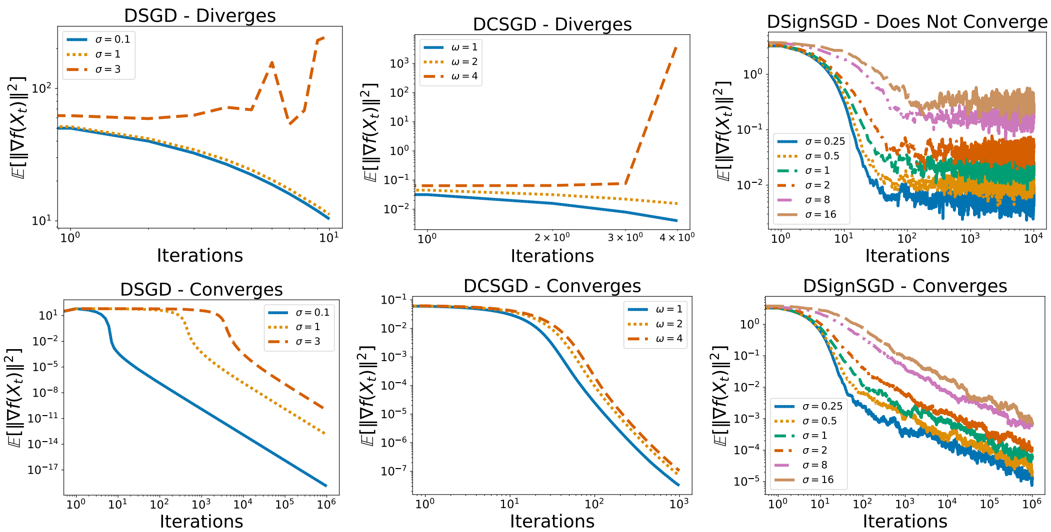

Figure 1: We optimize $f(x) = \frac{x^4}{4}$ with batch noise of variance $\sigma^2 \|\nabla f(x)\|_2^2$ for different values of σ: As per Thm. 4, DSGD diverges when σ is too large if normalization **is not employed** (Top-Left) but always converges if it **is employed** (Bottom-Left); We optimize $f(x) = \frac{\sum_{j=1}^{1000}(x_j)^4}{4}$ with batch noise of variance $\sigma^2 \|\nabla f(x)\|_2^2$ and use *Random Sparsification* for different compression rates ω: As per Thm. 5, DCSGD diverges if ω is too large, if normalization **is not employed** but it converges in all cases if it **is employed** (Bottom-Center); We optimize $f(x) = \frac{x^4}{4}$ with batch noise of **unbounded expected value** and for different *scale parameters* σ: As per Thm. 7, DSignSGD does not converge to 0 *without* a proper learning rate scheduler (Top-Right), but does converge *with* (Bottom-Right)

Intuition: This result showcases the crucial role of the gradient noise structure and the regularity of the loss landscape: If $\overline{\sigma_1^2} L_1 > 0$, normalizing the update step naturally emerges to ensure convergence. In particular, when $\overline{\sigma_1} = 0$, i.e., the noise has bounded variance, the standard stepsize schedule derived under L-smoothness remains effective under (L_0, L_1)-smoothness, without significant degradation in the convergence guarantee. According to our result, neither normalization nor gradient clipping is necessary to achieve this guarantee when $\overline{\sigma_1} = 0$. Interestingly, existing *upper bounds* obtained via *discrete-time* convergence analysis for standard SGD [21] exhibit worse dependence on L_0 and L_1 compared to the corresponding bounds for SGD *with* clipping [11, 18] or normalization [15]. These findings further motivate the investigation of the tightness of existing upper bounds in the discrete-time setting – a direction we leave open for future work.

Theorem 5 ***(DCSGD, unbiased compression, bounded variance)*** *Let f be (L_0, L_1)-smooth and each agent i have bounded variance σ_i^2, $\overline{\sigma^2} := \frac{1}{N}\sum_{i=1}^N \sigma_i^2$, and $\overline{\sigma^2 \omega} := \frac{1}{N}\sum_{i=1}^N \sigma_i^2 \omega_i$. For arbitrary $\epsilon \in (0,1)$, assume*

$$\eta \eta_t \leq \frac{2N\epsilon}{\overline{\omega} L_0 + d\left(\overline{\sigma^2} + \overline{\sigma^2 \omega}\right) L_1 + \overline{\omega} L_1 \mathbb{E}\left[\|\nabla f(X_t)\|_2\right]}. \tag{3}$$

Then, for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1(1-\epsilon)}\left(f(X_0) - f(X_*) + \phi_t^2 \frac{\eta(L_0+L_1)d\left(\overline{\sigma^2}+\overline{\sigma^2\omega}\right)}{2N}\right) \overset{t\to\infty}{\to} 0. \tag{4}$$

Intuition: This result showcases the crucial role of the gradient compression and the regularity of the loss landscape: If $L_1\overline{\omega} > 0$, normalizing the update step naturally emerges to ensure convergence. Additionally, one can draw a parallel between the normalization requirement for DSGD prescribed in Eq. 1 and that of DCSGD in Eq. 3: DCSGD with bounded variance σ^2 and compression rate ω is essentially equivalent to DSGD with (σ_0^2, σ_1^2)-variance where $\sigma_0^2 = \omega$ and $\sigma_1^2 = d(\sigma^2 + \omega)$.

One can generalize this result to cover the potentially unbounded variance setting.

Theorem 6 ***(DCSGD, unbiased compression, unbounded variance)*** *Let f be (L_0, L_1)-smooth, and each agent have $(\sigma_{0,i}^2, \sigma_{1,i}^2)$-variance. Define $\overline{\sigma_0^2} := \frac{1}{N}\sum_{i=1}^N \sigma_{0,i}^2$, $\overline{\sigma_1^2} := \frac{1}{N}\sum_{i=1}^N \sigma_{1,i}^2$, $\overline{\sigma_0^2\omega} := \frac{1}{N}\sum_{i=1}^N \sigma_{i,0}^2\omega_i$, and $\overline{\sigma_1^2\omega} := \frac{1}{N}\sum_{i=1}^N \sigma_{i,1}^2\omega_i$. For an arbitrary $\epsilon \in (0,1)$, assume*

$$\eta\eta_t \leq \frac{2N\epsilon}{L_0(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2})) + L_1 d\left(\overline{\sigma_0^2} + \overline{\sigma_0^2\omega}\right) + L_1(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2}))\mathbb{E}\left[\|\nabla f(X_t)\|_2\right]}. \tag{5}$$

Then, for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{(1-\epsilon)\phi_t^1}\left(f(X_0) - f(X_*) + \phi_t^2 \frac{L_0(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2})) + L_1 d\left(\overline{\sigma_0^2} + \overline{\sigma_0^2\omega}\right)}{2N}\right) \overset{t\to\infty}{\to} 0. \tag{6}$$

Intuition: This result showcases the crucial role of the gradient noise structure, the regularity of the loss landscape, and the compression scheme: If $L_1(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2})) > 0$, normalizing the update step naturally emerges to ensure convergence.

DSignSGD, structured noise, unbounded expected value. To provide tight results for the convergence of DSignSGD under **unbounded second and even first moments**, we additionally assume structured (heavy-tailed) noise following a student-t distribution: $\nabla f_{\gamma_i}(x) = \nabla f(x) + \sqrt{\Sigma_i}Z_i$ s.t. $Z_i \sim t_\nu(0, I_d)$, ν are the d.o.f, and *scale matrices*[1] $\Sigma_i = \text{diag}(\sigma_{1,i}^2, \cdots, \sigma_{d,i}^2)$. Note that if $\nu = 1$, the **expected value** of Z_i is **unbounded**, thus modeling much more pathological noise than simple (σ_0^2, σ_1^2)-variance.

Theorem 7 *Let f be (L_0, L_1)-smooth, $\eta\eta_t < \frac{2N\ell_\nu}{\sigma_{\mathcal{H},1}L_1 d}$, $\Sigma_i \leq \sigma_{max,i}^2$, $\sigma_{\mathcal{H},1}$ be the harmonic mean of $\{\sigma_{max,i}\}$, and $\ell_\nu > 0$ a constant. Then, for a random time $\tilde{t}$ with distribution $\frac{\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2\frac{\eta L_1 d}{2N}}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2\frac{\eta L_1 d}{2N}}$, we have that*

$$\mathbb{E}\|\nabla f(X_{\tilde{t}})\|_2^2 \leq \frac{1}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2\frac{\eta L_1 d}{2N}}\left(f(X_0) - f(X_*) + \frac{\eta(L_0 + L_1)d\phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0. \tag{7}$$

5. Conclusion

In this paper, we provided the first application of SDEs to (L_0, L_1)-smooth problems, deriving the first convergence guarantees for DSGD, DCSGD, and DSignSGD under such a condition as we coupled it with flexible batch noise assumptions. Importantly, we show that some sort of adaptivity is beneficial to ensure the convergence of stochastic optimizers. On one hand, an adaptive method such as DSignSGD converges even under heavy-tailed noise of **unbounded** expected value. On the other hand, for DCSGD, if either the compression rate $\overline{\omega}$ or the $\overline{\sigma_1^2}$ is positive, normalizing the updates emerges naturally as a strategy to ensure convergence. These findings prompt us to include the study of Normalized SGD under heavy-tailed noise in future work. Our final message is that the success of adaptive methods in Deep Learning has to be partially credited to the fact that their updates are, to some extent, normalized, thus actively countering the destabilizing effects of ill-conditioned landscapes even under large and possibly heavy-tailed noise.

1. These are *not* covariance matrices, but we use the same notation to facilitate comparability.

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

[2] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413. PMLR, 2018.

[3] 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*, 2018.

[4] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023.

[5] Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, and Aurelien Lucchi. An sde for modeling sam: Theory and insights. In *International Conference on Machine Learning*, pages 25209–25253. PMLR, 2023.

[6] Enea Monzio Compagnoni, Antonio Orvieto, Hans Kersting, Frank Proske, and Aurelien Lucchi. Sdes for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 4834–4842. PMLR, 2024.

[7] Enea Monzio Compagnoni, Rustem Islamov, Frank Norbert Proske, and Aurelien Lucchi. Unbiased and sign compression in distributed learning: Comparing noise resilience via SDEs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=RRrftHtEfK.

[8] Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ww3CLRhF1v.

[9] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35: 9955–9968, 2022.

[10] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 89–160. PMLR, 2023.

[11] Ofir Gaash, Kfir Yehuda Levy, and Yair Carmon. Convergence of clipped sgd on convex (L_0, L_1)-smooth functions. *arXiv preprint arXiv:2502.16492*, 2025.

[12] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

[13] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, page 1, 2021.

[14] Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. *International Conference on Learning Representations*, 2025.

[15] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR, 2024.

[16] Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *ICANN 2018*, 2018.

[17] Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under (l_0, l_1)-smoothness: Normalization and momentum. *arXiv preprint arXiv:2410.16871*, 2024.

[18] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

[19] Nikita Kornilov, Philip Zmushko, Andrei Semenov, Alexander Gasnikov, and Alexander Beznosikov. Sign operator for coping with heavy-tailed noise: High probability convergence bounds with extensions to distributed optimization and comparison oracle. *arXiv preprint arXiv:2502.07923*, 2025.

[20] Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

[21] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36: 40238–40271, 2023.

[22] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

[24] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1): 1474–1520, 2019.

[25] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[26] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, 2022.

[27] Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of SGD with gradient clipping in high-dimensions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jmN1zXMq0O.

[28] GN Mil'shtein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

[29] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, pages 1–16, 2024.

[30] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021.

[31] Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[32] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

[33] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[34] Mher Safaryan and Peter Richtarik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

[36] Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

[37] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

[38] Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *ArXiv*, abs/2101.12176, 2021.

[39] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

[40] Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. *International Conference on Learning Representations*, 2025.

[41] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

[42] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

[43] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[44] Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signsgd in high-dimensions: Quantifying preconditioning and noise-compression effects. *arXiv preprint arXiv:2411.12135*, 2024.

[45] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[46] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 2020.

[47] Jim Zhao, Aurelien Lucchi, Frank Norbert Proske, Antonio Orvieto, and Hans Kersting. Batch size selection by stochastic optimal control. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

[48] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

Appendix A. Theoretical Results

A.1. Distributed SGD

The following is the SDE model of DSGD (see Theorem 3.2 in [7]). Let us consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of

$$dX_t = -\nabla f(X_t)dt + \sqrt{\frac{\eta}{N}}\sqrt{\hat{\Sigma}(X_t)}dW_t, \tag{8}$$

where $\hat{\Sigma}(x) := \frac{1}{N}\sum_{i=1}^{N}\Sigma_i(x)$ is the average of the covariance matrices of the N agents.

Theorem 8 *Let f be (L_0, L_1)-smooth, $\|\Sigma_i(x)\|_\infty < \sigma_{0,i}^2 + \sigma_{1,i}^2\|\nabla f(x)\|_2^2$, the learning rate scheduler η_t s.t. $\phi_t^i = \int_0^t (\eta_s)^i ds$, $\phi_t^1 \overset{t\to\infty}{\to} \infty$, $\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0$, $\overline{\sigma_0^2} := \frac{1}{N}\sum_{i=1}^{N}\sigma_{0,i}^2$, and $\overline{\sigma_1^2} := \frac{1}{N}\sum_{i=1}^{N}\sigma_{1,i}^2$. Then, for $0 < \epsilon < 1$,*

$$\eta\eta_t < \frac{2N\epsilon}{d\left(\overline{\sigma_1^2}L_0 + \overline{\sigma_0^2}L_1 + L_1\overline{\sigma_1^2}\mathbb{E}\left[\|\nabla f(X_t)\|_2\right]\right)}, \tag{9}$$

and for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1(1-\epsilon)}\left(f(X_0) - f(X_*) + \phi_t^2\frac{\eta d(L_0+L_1)(\overline{\sigma_0^2}+\overline{\sigma_1^2})}{2N}\right) \overset{t\to\infty}{\to} 0. \tag{10}$$

Proof Using Itô's Lemma and using a learning rate scheduler η_t during the derivation of the SDE, we have

$$d(f(X_t) - f(X_*)) = -\eta_t\|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) + (\eta_t)^2\frac{\eta}{2N}\text{Tr}(\nabla^2 f(X_t)\tilde{\Sigma}(X_t))dt \tag{11}$$

$$\leq -\eta_t\|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) \tag{12}$$

$$+ (\eta_t)^2\frac{\eta(\overline{\sigma_0^2} + \overline{\sigma_1^2}\|\nabla f(X_t)\|_2^2)d(L_0 + L_1\|\nabla f(X_t)\|)}{2N}dt. \tag{13}$$

Phase 1: If $\|\nabla f(X_t)\| \leq 1$, then the proof and conditions are the same as the L-smoothness case. Let us observe that since $\int_0^t \frac{\eta_s}{\phi_t^1}ds = 1$, the function $s \mapsto \frac{\eta_s}{\phi_t^1}$ defines a probability distribution and let $\tilde{t}$ have that distribution. Then, by integrating over time and by the Law of the Unconscious Statistician, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] = \frac{1}{\phi_t^1}\int_0^t \|\nabla f(X_s)\|_2^2\eta_s ds, \tag{14}$$

meaning that

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \leq \frac{f(X_0) - f(X_*)}{\phi_t^1} + \frac{\eta(L_0+L_1)(\overline{\sigma_0^2}+\overline{\sigma_1^2})d}{2N}\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0. \tag{15}$$

Phase 2: If $\|\nabla f(X_t)\| > 1$, we have

$$d(f(X_t) - f(X_*)) = -\eta_t \|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) + (\eta_t)^2 \frac{\eta}{2N} \text{Tr}(\nabla^2 f(X_t) \tilde{\Sigma}(X_t)) dt \tag{16}$$

$$\leq -\eta_t \|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) \tag{17}$$

$$+ (\eta_t)^2 \frac{\eta(\overline{\sigma_0^2} + \overline{\sigma_1^2} \|\nabla f(X_t)\|_2^2) d(L_0 + L_1 \|\nabla f(X_t)\|)}{2N} dt \tag{18}$$

$$= -\eta_t \|\nabla f(X_t)\|_2^2 \left(1 - \frac{\eta_t \eta \overline{\sigma^2} d}{2N} \left(\overline{\sigma_1^2} L_0 + \overline{\sigma_0^2} L_1 + L_1 \overline{\sigma_1^2} \|\nabla f(X_t)\|_2\right)\right) \tag{19}$$

$$+ (\eta_t)^2 \frac{\eta \overline{\sigma_0^2} d L_0}{2N} \tag{20}$$

Therefore, for $0 < \epsilon < 1$ we have that if

$$\eta \eta_t < \frac{2N\epsilon}{d\left(\overline{\sigma_1^2} L_0 + \overline{\sigma_0^2} L_1 + L_1 \overline{\sigma_1^2} \|\nabla f(X_t)\|_2\right)}, \tag{21}$$

and therefore that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1 (1-\epsilon)} \left(f(X_0) - f(X_*) + \phi_t^2 \frac{\eta L_0 d \overline{\sigma^2}}{2N}\right) \overset{t \to \infty}{\to} 0, \tag{22}$$

where $\hat{t}$, is a random time with distribution $\frac{\eta_{\hat{t}}}{\phi_t^1}$.

By taking a worst-case scenario approach, we merge these two bounds into a single one:

$$d(f(X_t) - f(X_*)) \leq -\eta_t \|\nabla f(X_t)\|_2^2 \left(1 - \frac{\eta_t \eta \overline{\sigma^2} d}{2N} \left(\overline{\sigma_1^2} L_0 + \overline{\sigma_0^2} L_1 + L_1 \overline{\sigma_1^2} \|\nabla f(X_t)\|_2\right)\right) \tag{23}$$

$$+ (\eta_t)^2 \frac{\eta d (L_0 + L_1)(\overline{\sigma_0^2} + \overline{\sigma_1^2})}{2N}, \tag{24}$$

and, therefore, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1 (1-\epsilon)} \left(f(X_0) - f(X_*) + \phi_t^2 \frac{\eta d (L_0 + L_1)(\overline{\sigma_0^2} + \overline{\sigma_1^2})}{2N}\right) \overset{t \to \infty}{\to} 0, \tag{25}$$

where $\hat{t}$, is a random time with distribution $\frac{\eta_{\hat{t}}}{\phi_t^1}$.

Finally, for practical reasons, we leverage the distributed setting to tighten the requirements on the learning rate scheduler to make it experimentally viable, and rather require

$$\eta \eta_t < \frac{2N\epsilon}{d\left(\overline{\sigma_1^2} L_0 + \overline{\sigma_0^2} L_1 + L_1 \overline{\sigma_1^2} \mathbb{E}\left[\|\nabla f(X_t)\|_2\right)\right]}. \tag{26}$$

■

A.2. Distributed Compressed SGD with Unbiased Compression

The following is the SDE model of DCSGD (see Theorem 3.6 in [7]). Let us consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of

$$dX_t = -\nabla f(X_t)dt + \sqrt{\frac{\eta}{N}}\sqrt{\tilde{\Sigma}(X_t)}dW_t, \tag{27}$$

where for $\Phi_{\xi_i,\gamma_i}(x) := \mathcal{C}_{\xi_i}\left(\nabla f_{\gamma_i}(x)\right) - \nabla f_{\gamma_i}(x)$

$$\tilde{\Sigma}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\xi_i\gamma_i}\left[\Phi_{\xi_i,\gamma_i}(x)\Phi_{\xi_i,\gamma_i}(x)^\top\right] + \Sigma_i(x)\right). \tag{28}$$

Theorem 9 *Let f be (L_0, L_1)-smooth, the learning rate scheduler η_t such that $\phi_t^i = \int_0^t (\eta_s)^i ds$, $\phi_t^1 \overset{t\to\infty}{\to} \infty$, $\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0$, and $\overline{\sigma^2\omega} := \frac{1}{N}\sum_{i=1}^N \sigma_i^2\omega_i$. Then, for $0 < \epsilon < 1$,*

$$\eta\eta_t < \frac{2N\epsilon}{\overline{\omega}L_0 + \left(\overline{\sigma^2}d + d\overline{\sigma^2\omega}\right)L_1 + \overline{\omega}L_1\mathbb{E}\left[\|\nabla f(X_t)\|_2\right]}, \tag{29}$$

and for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \le \frac{1}{\phi_t^1(1-\epsilon)}\left(f(X_0) - f(X_*) + \phi_t^2\frac{\eta(L_0+L_1)d\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)}{2N}\right) \overset{t\to\infty}{\to} 0. \tag{30}$$

Proof Since it holds that

$$\mathbb{E}_{\xi_i,\gamma_i}\|(\mathcal{C}_{\xi_i}\left(\nabla f_{\gamma_i}(x)\right) - \nabla f(x))\|_2^2 \le \omega_i\|\nabla f(x)\|_2^2 + d\sigma_i^2(\omega_i+1),$$

we have that

$$d(f(X_t) - f(X_*)) = -\eta_t\|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) \tag{31}$$

$$+ (\eta_t)^2\frac{\eta(L_0 + L_1\|\nabla f(X_t)\|_2)}{2N}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\xi_i,\gamma_i}\|(\mathcal{C}_{\xi_i}\left(\nabla f_{\gamma_i}(x)\right) - \nabla f(x))\|_2^2\right)dt \tag{32}$$

$$\le -\eta_t\|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) \tag{33}$$

$$+ (\eta_t)^2\frac{\eta(L_0 + L_1\|\nabla f(X_t)\|_2)}{2N}\left(\overline{\omega}\|\nabla f(X_t)\|_2^2 + \overline{\sigma^2}d + d\overline{\sigma^2\omega}\right)dt \tag{34}$$

Phase 1: If $\|\nabla f(X_t)\|_2 \le 1$, then we have that

$$\mathbb{E}\left[\|\nabla f(X_t)\|_2^2\right]\left(\eta_t - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}(\eta_t)^2\right)dt \le -d(f(X_t) - f(X_*)) \tag{35}$$

$$+ (\eta_t)^2\frac{\eta(L_0+L_1)d}{2N}\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)dt. \tag{36}$$

Let us now observe that since $\int_0^t \frac{\eta_s - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\eta_s^2}{\phi_t^1 - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\phi_t^2}ds = 1$, the function $s \mapsto \frac{\eta_s - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\eta_s^2}{\phi_t^1 - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\phi_t^2}$ defines a probability distribution and let $\tilde{t}$ have that distribution. Then by integrating over time and by the Law of the Unconscious Statistician, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] = \frac{1}{\phi_t^1 - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\phi_t^2}\int_0^t \|\nabla f(X_s)\|_2^2\left(\eta_s - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\eta_s^2\right)ds, \tag{37}$$

meaning that

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1 - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\phi_t^2}\left(f(X_0) - f(X_*) + \phi_t^2\frac{\eta(L_0+L_1)d}{2N}\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)\right) \overset{t\to\infty}{\to} 0, \quad (38)$$

where $\tilde{t}$, is a random time with distribution $\frac{\eta_{\tilde{t}} - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}(\eta_{\tilde{t}})^2}{\phi_t^1 - \frac{\eta(L_0+L_1)\overline{\omega}}{2N}\phi_t^2}$.

Phase 2: If $\|\nabla f(X_t)\|_2 > 1$, we have that

$$\begin{aligned} d(f(X_t) - f(X_*)) &\leq -\eta_t\|\nabla f(X_t)\|_2^2 dt + \mathcal{O}(\text{Noise}) && (39)\\ &+ (\eta_t)^2\frac{\eta(L_0 + L_1\|\nabla f(X_t)\|_2)}{2N}\left(\overline{\omega}\|\nabla f(X_t)\|_2^2 + \overline{\sigma^2}d + d\overline{\sigma^2\omega}\right)dt && (40)\\ &\leq -\eta_t\|\nabla f(X_t)\|_2^2\left(1 - \frac{\eta_t\eta}{2N}\left(\overline{\omega}L_0 + d\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)L_1 + \overline{\omega}L_1\|\nabla f(X_t)\|_2\right)\right) && (41)\\ &+ \eta_t^2\frac{\eta L_0 d}{2N}\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right). && (42)\end{aligned}$$

Therefore, for $0 < \epsilon < 1$ we have that if

$$\eta\eta_t < \frac{2N\epsilon}{\overline{\omega}L_0 + d\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)L_1 + \overline{\omega}L_1\|\nabla f(X_t)\|_2}, \quad (43)$$

then,

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1(1-\epsilon)}\left(f(X_0) - f(X_*) + \phi_t^2\frac{\eta L_0 d}{2N}\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)\right) \overset{t\to\infty}{\to} 0, \quad (44)$$

where $\hat{t}$, is a random time with distribution $\frac{\eta_{\hat{t}}}{\phi_t^1}$. Finally, for practical reasons, we leverage the distributed setting to tighten the requirements on the learning rate scheduler to make it experimentally viable, and rather require

$$\eta\eta_t < \frac{2N\epsilon}{\overline{\omega}L_0 + \left(\overline{\sigma^2}d + d\overline{\sigma^2\omega}\right)L_1 + \overline{\omega}L_1\mathbb{E}\left[\|\nabla f(X_t)\|_2\right]}, \quad (45)$$

By taking a worst-case scenario approach, we merge these two bounds into a single one and have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{\phi_t^1(1-\epsilon)}\left(f(X_0) - f(X_*) + \phi_t^2\frac{\eta(L_0+L_1)d\left(\overline{\sigma^2} + \overline{\sigma^2\omega}\right)}{2N}\right) \overset{t\to\infty}{\to} 0, \quad (46)$$

where $\hat{t}$, is a random time with distribution $\frac{\eta_{\hat{t}}}{\phi_t^1}$. ∎

Finally, one can generalize this result to cover the (σ_0^2, σ_1^2)-Variance.

Theorem 10 *Let f be (L_0, L_1)-smooth,* $\max(\Sigma_i(x)) < \sigma_{i,0}^2 + \sigma_{i,1}^2\|\nabla f(x)\|_2^2$*, the learning rate scheduler η_t such that $\phi_t^i = \int_0^t(\eta_s)^i ds$, $\phi_t^1 \overset{t\to\infty}{\to} \infty$, $\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0$, $\overline{\sigma_0^2} := \frac{1}{N}\sum_{i=1}^N \sigma_{0,i}^2$, $\overline{\sigma_1^2} := \frac{1}{N}\sum_{i=1}^N \sigma_{1,i}^2$, $\overline{\sigma_0^2\omega} := \frac{1}{N}\sum_{i=1}^N \sigma_{i,0}^2\omega_i$, and $\overline{\sigma_1^2\omega} := \frac{1}{N}\sum_{i=1}^N \sigma_{i,1}^2\omega_i$. Then, for $0 < \epsilon < 1$,*

$$\eta\eta_t < \frac{2N\epsilon}{L_0(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2})) + L_1 d\left(\overline{\sigma_0^2} + \overline{\sigma_0^2\omega}\right) + L_1(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2}))\mathbb{E}\left[\|\nabla f(X_t)\|_2\right]}, \quad (47)$$

and for a random time $\hat{t}$ with distribution $\frac{\eta_t}{\phi_t^1}$, we have that

$$\mathbb{E}\left[\|\nabla f(X_{\hat{t}})\|_2^2\right] \leq \frac{1}{(1-\epsilon)\phi_t^1}\left(f(X_0) - f(X_*) + \phi_t^2 \frac{L_0(\overline{\omega} + d(\overline{\sigma_1^2\omega} + \overline{\sigma_1^2})) + L_1 d\left(\overline{\sigma_0^2} + \overline{\sigma_0^2\omega}\right)}{2N}\right) \overset{t\to\infty}{\to} 0. \quad (48)$$

A.3. Distributed SignSGD

The following is the SDE model of DSignSGD (see Theorem 3.10 in [7]). Let us consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of

$$dX_t = -\frac{1}{N}\sum_{i=1}^{N}(1 - 2\mathbb{P}(\nabla f_{\gamma_i}(X_t) < 0))\, dt + \sqrt{\frac{\eta}{N}}\sqrt{\overline{\Sigma}(X_t)}dW_t. \quad (49)$$

where

$$\overline{\Sigma}(X_t) := \frac{1}{N}\sum_{i=1}^{N}\overline{\Sigma_i}(X_t), \quad (50)$$

and $\overline{\Sigma_i}(x) = \mathbb{E}[\xi_{\gamma_i}(x)\xi_{\gamma_i}(x)^\top]$ where $\xi_{\gamma_i}(x) := \text{sign}(\nabla f_{\gamma_i}(x)) - 1 + 2\mathbb{P}(\nabla f_{\gamma_i}(x) < 0)$ the noise in the sample sign $(\nabla f_{\gamma_i}(x))$.

Corollary 11 (Corollary C.10 in [7]) *If the stochastic gradients are $\nabla f_{\gamma_i}(x) = \nabla f(x) + \sqrt{\Sigma_i} Z_i$ such that $Z_i \sim t_\nu(0, I_d)$ does not depend on x, ν are the degrees of freedom, and scale matrices $\Sigma_i = \text{diag}(\sigma_{1,i}^2, \cdots, \sigma_{d,i}^2)$. Then, the SDE of DSignSGD is*

$$dX_t = -\frac{2}{N}\sum_{i=1}^{N}\Xi_\nu\left(\Sigma_i^{-\frac{1}{2}}\nabla f(X_t)\right)dt + \sqrt{\frac{\eta}{N}}\sqrt{\tilde{\Sigma}(X_t)}dW_t. \quad (51)$$

where $\Xi_\nu(x)$ is defined as $\Xi_\nu(x) := x\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)$, ${}_2F_1(a, b; c; x)$ is the hypergeometric function, and

$$\tilde{\Sigma}(X_t) := I_d - \frac{4}{N}\sum_{i=1}^{N}\left(\Xi_\nu\left(\Sigma_i^{-\frac{1}{2}}\nabla f(X_t)\right)\right)^2. \quad (52)$$

Following their notation, we split the analysis of DSignSGD into **Phases** of the dynamics depending on the value of the signal-to-ratio $Y_t^i := \sqrt{B}\Sigma_i^{-\frac{1}{2}}\nabla f(X_t)$ and a constant ψ_ν:

- **Phase 1:** If $\left|Y_t^i\right| > \psi_\nu$;
- **Phase 2:** $1 < \left|Y_t^i\right| < \psi_\nu$;
- **Phase 3:** $\left|Y_t^i\right| < 1$.

In the following, the constants $\mathbf{q}_\nu^+$, $\mathbf{q}_\nu^-$, $\hat{q}_\nu$, m_ν, ℓ_ν, and ψ_ν are defined in Proposition C.11 [7].

Theorem 12 *Let f be (L_0, L_1)-smooth, η_t a learning rate scheduler such that $\phi_t^i = \int_0^t (\eta_s)^i ds$, $\phi_t^1 \overset{t\to\infty}{\to} \infty$, $\frac{\phi_t^2}{\phi_t^1} \overset{t\to\infty}{\to} 0$, $\Sigma_i \leq \sigma_{max,i}^2$, $\sigma_{\mathcal{H},1}$ be the harmonic mean of $\{\sigma_{max,i}\}$, and $\ell_\nu > 0$ a constant. Then, for a random time $\tilde{t}$ with distribution $\frac{\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2\frac{\eta L_1 d}{2N}}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2\frac{\eta L_1 d}{2N}}$, we have that*

$$\mathbb{E}\|\nabla f(X_{\tilde{t}})\|_2^2 \leq \frac{1}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2\frac{\eta L_1 d}{2N}}\left(f(X_0) - f(X_*) + \frac{\eta(L_0 + L_1)d\phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0. \quad (53)$$

In particular:

In Phase 1,

$$\mathbb{E}\left[\|\nabla f\left(X_{\tilde{t}^1}\right)\|_1\right] \leq \frac{f(X_0) - f(X_*)}{\phi_t^1} \overset{t\to\infty}{\to} 0, \tag{54}$$

where $\tilde{t}^1$ is a random time with distribution $\frac{\eta_t}{\phi_t^1}$;

In Phase 2:

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,1)}}\right)\|_2^2 + \frac{\hat{q}_\nu \phi_t^1}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,2)}}\right)\|_1 \leq \frac{f(X_0) - f(X_*)}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \tag{55}$$

$$+ \frac{\phi_t^2 \frac{\eta(L_0+L_1)d}{2N}}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \overset{t\to\infty}{\to} 0, \tag{56}$$

where $\tilde{t}^{(2,1)}$ and $\tilde{t}^{(2,2)}$ are a random times with distributions $\frac{\eta_t m_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2 \frac{\eta L_1 d}{2N}}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}}$ and $\frac{\eta_t}{\phi_t^1}$, respectively.

In Phase 3:

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^3}\right)\|_2^2 \leq \frac{1}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \left(f(X_0) - f(X_*) + \frac{\eta(L_0 + L_1) d \phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0, \tag{57}$$

where $\tilde{t}^3$ is a random time with distribution $\frac{\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2 \frac{\eta L_1 d}{2N}}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}}$.

Proof Let us prove the above phase by phase:

For Phase 1,

$$d(f(X_t) - f(X_*)) = -\eta_t \nabla f(X_t) \operatorname{sign}(\nabla f(X_t)) dt = -\eta_t \|\nabla f(X_t)\|_1 dt \tag{58}$$

Therefore, by integrating over time, dividing by ϕ_t^1, and using the law of the unconscious statistician

$$\mathbb{E}\left[\|\nabla f\left(X_{\tilde{t}^1}\right)\|_1\right] \leq \frac{f(X_0) - f(X_*)}{\phi_t^1} \overset{t\to\infty}{\to} 0; \tag{59}$$

where $\tilde{t}^1$ is a random time with distribution $\frac{\eta_{\tilde{t}^1}}{\phi_t^1}$; ■

For Phase 2, using Itô on f, we have that

$$d(f(X_t) - f(X_*)) \leq -\eta_t m_\nu \sigma_{\mathcal{H},1}^{-1} \|\nabla f(X_t)\|_2^2 dt - \eta_t \hat{q}_\nu \|\nabla f(X_t)\|_1 dt \tag{60}$$

$$+ \eta_t^2 \frac{\eta(L_0 + L_1 \|\nabla f(X_t)\|_2) d}{2N} dt + \mathcal{O}(\text{Noise}) \tag{61}$$

Phase 2.1: if $\|\nabla f(X_t)\|_2 \leq 1$: By integrating over time and using the law of the unconscious statistician, we have

$$m_\nu \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(1,2)}}\right)\|_2^2 + \hat{q}_\nu \sigma_{\mathcal{H},1} \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,2)}}\right)\|_1 \leq \frac{\sigma_{\mathcal{H},1}}{\phi_t^1} \left(f(X_0) - f(X_*) + \frac{\eta(L_0 + L_1) d \phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0, \tag{62}$$

where $\tilde{t}^{(1,2)}$ and $\tilde{t}^{(2,2)}$ are random times with distribution $\frac{\eta_t}{\phi_t^1}$.

Phase 2.2: if $\|\nabla f(X_t)\|_2 > 1$: using Itô on f, we have that

$$\eta_t \left(m_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t \frac{\eta L_1 d}{2N} \right) \|\nabla f(X_t)\|_2^2 dt + \eta_t \hat{q}_\nu \|\nabla f(X_t)\|_1 dt \leq -d(f(X_t) - f(X_*)) \tag{63}$$

$$+ \eta_t^2 \frac{\eta L_0 d}{2N} dt + \mathcal{O}(\text{Noise}). \tag{64}$$

Therefore, integrating over time and then dividing by $\left(\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N} \right) (\phi_t^1)$,

$$\frac{\int_0^t \eta_s \left(m_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_s \frac{\eta L_1 d}{2N} \right) \|\nabla f(X_s)\|_2^2 ds + \eta_s \hat{q}_\nu \|\nabla f(X_s)\|_1 ds}{\left(\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N} \right) (\phi_t^1)} \leq \frac{f(X_0) - f(X_*)}{\left(\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N} \right) (\phi_t^1)} \tag{65}$$

$$+ \frac{\int_0^t \eta_s^2 \frac{\eta L_0 d}{2N} ds}{\left(\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N} \right) (\phi_t^1)} + \mathcal{O}(\text{Noise})$$

and using the law of the unconscious statistician as $\eta_t < \frac{2N m_\nu}{\sigma_{\mathcal{H},1} \eta L_1 d}$, and multiplying by ϕ_t^1, we have

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,1)}}\right)\|_2^2 + \frac{\hat{q}_\nu \phi_t^1}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,2)}}\right)\|_1 \leq \frac{f(X_0) - f(X_*)}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \tag{66}$$

$$+ \frac{\phi_t^2 \frac{\eta L_0 d}{2N} dt}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \overset{t \to \infty}{\to} 0 \tag{67}$$

By taking a worst-case scenario approach, we merge these two bounds into a single one:

$$\eta_t \left(m_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t \frac{\eta L_1 d}{2N} \right) \|\nabla f(X_t)\|_2^2 dt + \eta_t \hat{q}_\nu \|\nabla f(X_t)\|_1 dt \leq -d(f(X_t) - f(X_*)) \tag{68}$$

$$+ \eta_t^2 \frac{\eta (L_0 + L_1) d}{2N} dt + \mathcal{O}(\text{Noise}), \tag{69}$$

and therefore that

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,1)}}\right)\|_2^2 + \frac{\hat{q}_\nu \phi_t^1}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(2,2)}}\right)\|_1 \leq \frac{f(X_0) - f(X_*)}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \tag{70}$$

$$+ \frac{\phi_t^2 \frac{\eta (L_0 + L_1) d}{2N} dt}{\phi_t^1 m_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \overset{t \to \infty}{\to} 0 \tag{71}$$

For Phase 3, using Itô on f, we have that

$$d(f(X_t) - f(X_*)) \leq -\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} \|\nabla f(X_t)\|_2^2 dt + \eta_t^2 \frac{\eta (L_0 + L_1 \|\nabla f(X_t)\|_2) d}{2N} dt + \mathcal{O}(\text{Noise}) \tag{72}$$

Phase 3.1: if $\|\nabla f(X_t)\|_2 \leq 1$: By integrating over time and using the law of the unconscious statistician, we have

$$\ell_\nu \mathbb{E}\|\nabla f\left(X_{\tilde{t}^{(1,2)}}\right)\|_2^2 \leq \frac{\sigma_{\mathcal{H},1}}{\phi_t^1} \left(f(X_0) - f(X_*) + \frac{\eta (L_0 + L_1) d \phi_t^2}{2N} \right) \overset{t \to \infty}{\to} 0, \tag{73}$$

where $\tilde{t}^{(1,2)}$ is a random time with distribution $\frac{\eta_t}{\phi_t^1}$.

Phase 3.2: if $\|\nabla f(X_t)\|_2 > 1$: using Itô on f, we have that

$$\eta_t \left(\ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t \frac{\eta L_1 d}{2N}\right) \|\nabla f(X_t)\|_2^2 dt \leq -d(f(X_t) - f(X_*)) + \eta_t^2 \frac{\eta L_0 d}{2N} dt + \mathcal{O}(\text{Noise}). \tag{74}$$

Therefore, integrating over time and using the law of the unconscious statistician as $\eta_t < \frac{2N\ell_\nu}{\sigma_{\mathcal{H},1}\eta L_1 d}$, we have

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^3}\right)\|_2^2 \leq \frac{1}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \left(f(X_0) - f(X_*) + \frac{\eta L_0 d \phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0, \tag{75}$$

where $\tilde{t}^3$ is a random time with distribution $\frac{\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2 \frac{\eta L_1 d}{2N}}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}}$.

By taking a worst-case scenario approach, we merge these two bounds into a single one:

$$\mathbb{E}\|\nabla f\left(X_{\tilde{t}^3}\right)\|_2^2 \leq \frac{1}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}} \left(f(X_0) - f(X_*) + \frac{\eta (L_0 + L_1) d \phi_t^2}{2N}\right) \overset{t\to\infty}{\to} 0, \tag{76}$$

where $\tilde{t}^3$ is a random time with distribution $\frac{\eta_t \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \eta_t^2 \frac{\eta L_1 d}{2N}}{\phi_t^1 \ell_\nu \sigma_{\mathcal{H},1}^{-1} - \phi_t^2 \frac{\eta L_1 d}{2N}}$.

A.4. Limitations

As noted by [25], the approximation power of SDEs can fail when the stepsize η is large or if certain conditions on ∇f and the noise covariance matrix are not met. Although these issues can be addressed by increasing the order of the weak approximation, we believe that the primary purpose of SDEs is to serve as simplification tools that enhance our intuition: We would not benefit significantly from added complexity.

Importantly, extensive experimental design empirically validated that the SDEs do track their respective optimizers precisely on a variety of architectures, including MLPs, CNNs, ResNets, and ViTs, [6–8, 30].

Appendix B. Experiments

B.1. DSGD - Figure 1 - (Left Column)

We optimize $f(x) = \frac{x^4}{4}$ as we inject gaussian noise with mean 0 and variance $\sigma^2 \|\nabla f(x)\|_2^2$ on the gradient. The learning rate is $\eta = 0.01$, $\sigma \in \{0.1, 1, 3\}$, and we average over 1000 runs. In the top figure, we use no scheduler, while in the bottom one we use a scheduler as per Eq. 1.

B.2. DCSGD - Figure 1 - (Center Column)

We optimize $f(x) = \frac{\sum_{j=1}^{1000}(x_j)^4}{4}$ as we inject gaussian noise with mean 0 and variance $\sigma^2 \|\nabla f(x)\|_2^2$ on the gradient. The learning rate is $\eta = 0.1$, $\sigma = 0.1\}$, we use *random sparsification* with $\omega \in \{1, 2, 4\}$, and we average over 1000 runs. In the top figure, we use no scheduler, while in the bottom one we use a scheduler as per Eq. 3.

B.3. DSignSGD - Figure 1 - (Right Column)

We optimize $f(x) = \frac{x^4}{4}$ as we inject student's t noise with $\nu = 1$ and scale parameters σ on the gradient. The learning rate is $\eta = 0.1$, $\sigma \in \{0.25, 0.5, 1, 2, 8, 16\}$, and we average over 10000 runs. In the top figure, we use no scheduler, while in the bottom one we use a scheduler as per Theorem 7, e.g. $\eta_t = \frac{1}{\sqrt{t+1}}$.