Journal of Machine Learning Research 25 (2024) 1-36 Submitted 4/23; Revised 7/24; Published 8/24

High Probability Convergence Bounds for Non-convex Stochastic Gradient Descent with Sub-Weibull Noise

Liam Madden LIAM@ECE.UBC.CA
Department of Electrical and Computer Engineering
University of British Columbia

Emiliano Dall'Anese EDALLANE@BU.EDU
Department of Electrical and Computer Engineering
Boston University

Stephen Becker STEPHEN.BECKER@COLORADO.EDU
Department of Applied Mathematics
University of Colorado Boulder

Editor: Simon Lacoste-Julien

Abstract

Stochastic gradient descent is one of the most common iterative algorithms used in machine learning and its convergence analysis is a rich area of research. Understanding its convergence properties can help inform what modifications of it to use in different settings. However, most theoretical results either assume convexity or only provide convergence results in mean. This paper, on the other hand, proves convergence bounds in high probability without assuming convexity. Assuming strong smoothness, we prove high probability convergence bounds in two settings: (1) assuming the Polyak-Łojasiewicz inequality and norm sub-Gaussian gradient noise and (2) assuming norm sub-Weibull gradient noise. In the second setting, as an intermediate step to proving convergence, we prove a sub-Weibull martingale difference sequence self-normalized concentration inequality of independent interest. It extends Freedman-type concentration beyond the sub-exponential threshold to heavier-tailed martingale difference sequences. We also provide a post-processing method that picks a single iterate with a provable convergence guarantee as opposed to the usual bound for the unknown best iterate. Our convergence result for sub-Weibull noise extends the regime where stochastic gradient descent has equal or better convergence guarantees than stochastic gradient descent with modifications such as clipping, momentum, and normalization.

Keywords: stochastic gradient descent, convergence bounds, sub-Weibull distributions, Polyak-Łojasiewicz inequality, Freedman inequality

1. Introduction

Stochastic gradient descent (SGD) and its variants are some of the most commonly used algorithms in machine learning. In particular, they are used for training neural network and transformer models, models that have achieved considerable success on image classification and language processing tasks in recent years. The training in this case is non-convex and smooth for many activation/attention functions, such as sigmoid, GELU, and softmax. Even for ReLU, which is not differentiable, the training is smooth over most of the parameter

space and avoidance of the non-smooth part is dealt with separately. Thus, the smooth non-convex convergence analysis of SGD has far-reaching influence on the field of machine learning.

There is a large literature on the almost sure and mean convergence of SGD assuming strong smoothness. Bertsekas and Tsitsiklis (2000) prove that SGD almost surely converges to a first-order stationary point assuming strong smoothness and the relaxed growth noise condition; more recently, Patel (2021) showed the same result with a different proof technique. Ghadimi and Lan (2013) prove that the mean of the squared gradient norm of SGD converges to zero at a rate of $O(1/\sqrt{T})$ assuming strong smoothness and the bounded variance noise condition, where T is the number of SGD iterations. Khaled and Richtárik (2023) prove the same but with the expected smoothness noise condition. Sebbouh et al. (2021) strengthen this to an almost sure convergence rate. Assuming strong smoothness and convexity, the tight mean convergence rate of the squared gradient norm of SGD is $O(1/\sqrt{T})$ (Nemirovsky and Yudin, 1983, Thm. 5.3.1). In fact, this is the optimal rate for all stochastic first-order methods assuming only strong smoothness (Arjevani et al., 2023). Thus, the $O(1/\sqrt{T})$ mean convergence rate of Ghadimi and Lan (2013) is tight and shows that SGD is optimal in the smooth non-convex setting.

However, mean convergence is not the end of the story. A convergence guarantee is generally required with some arbitrary probability, $1-\delta$. For a single run of SGD, using Markov's inequality gives a $1/\delta$ scaling to the bound. If one can re-run SGD many times (say, r times), then by choosing the best run, δ can be relatively large since the overall success is at least by $1-\delta^r$. On the other hand, if just a single run of SGD is allowed, then the $1/\delta$ factor leads to nearly vacuous results, and hence a direct high probability convergence analysis is necessary to understand the behavior of unique runs of SGD.

Li and Orabona (2020) prove a $O(\log(T)\log(T/\delta)/\sqrt{T})$ high probability convergence rate for a weighted average of the squared gradient norms of SGD assuming strong smoothness and norm sub-Gaussian noise. However, a series of recent papers (Gürbüzbalaban et al., 2021; Şimşekli et al., 2019; Panigrahi et al., 2019) suggest that norm sub-Gaussian noise is often not satisfied. While the central limit theorem can be used to heuristically justify norm sub-Gaussian noise for mini-batch SGD with large batch-sizes, it cannot for small batch-sizes. In particular, Panigrahi et al. (2019) provide examples where the noise is Gaussian for a batch-size of 4096, is not Gaussian for a batch-size of 32, and starts out Gaussian then becomes non-Gaussian for a batch-size of 256. Gürbüzbalaban et al. (2021) and Şimşekli et al. (2019) suggest instead assuming that the pth moment of the norm of the noise is bounded, where $p \in (1,2)$. Scaman and Malherbe (2020) prove a $O(1/T^{(p-1)/p}/\delta^{(2+p)/(2p)})$ high probability convergence rate for SGD assuming strong smoothness and bounded pth moment noise for $p \in (1,2)$. By allowing noise with possibly infinite variance, SGD converges at a slower rate in terms of T. Moreover, the dependence on $1/\delta$ is polynomial rather than logarithmic. Cutkosky and Mehta (2021) prove a $O(\log(T/\delta)/T^{(p-1)/(3p-2)})$ high probability convergence rate for clipped SGD (which uses clipping, momentum, and normalization) assuming strong smoothness and bounded pth moment noise for $p \in (1,2)$, thus improving the dependence on $1/\delta$ to logarithmic, but clipped SGD will still have a slower convergence rate in terms of T even when the noise is norm sub-Gaussian. So, knowledge about the type of noise should actually affect whether we use SGD or clipped SGD. This motivates

the question: **How heavy can the noise be for SGD to have a high probability convergence rate that depends logarithmically on $1/\delta$?**

Towards this end, we consider norm sub-Weibull noise. While it is much lighter tailed than bounded pth moment noise for $p \in (1, 2)$, it is a natural extension of norm sub-Gaussian noise and it turns out that it admits a convergence rate for SGD with logarithmic dependence on $1/\delta$. This leads us to the first contribution of our paper. **Assuming strong smoothness and norm sub-Weibull noise, we prove a $O((\log(T)\log(1/\delta)^{2\theta} + \log(T/\delta)^{\min\{0,\theta-1\}}\log(1/\delta))/\sqrt{T})$ high probability convergence rate for a weighted average of the squared gradient norms of SGD, where θ is the sub-Weibull tail weight.** This is given in Theorem 15. In the special case of norm sub-Gaussian noise, which is $\theta = 1/2$, this becomes $O(\log(T)\log(1/\delta)/\sqrt{T})$, which slightly improves the result of Li and Orabona (2020). For more general $\theta > 1/2$, we make the additional assumption that the objective function is Lipschitz continuous. At its most basic, the Lipschitz continuity assumption says that the norm of the true gradient is bounded for all of the iterates. Li and Liu (2022), building off of our pre-print, relax this to only assuming that the step-size times the true gradient is bounded for all of the iterates, which is a weaker assumption since the step-size goes to zero.

However, hidden within these convergence rates is a subtle issue that requires a post-processing algorithm. All of these high probability convergence rates are for a weighted average of the squared gradient norms of the iterates rather than being for the squared gradient norm of a single iterate. While this implies a high probability convergence rate for the iterate with the smallest squared gradient norm, to actually determine which iterate has the smallest squared gradient norm, we would need to estimate the true gradient for each iterate! This leads us to another question: **Is there a post-processing strategy for a single run of SGD that is as efficient as 2-RSG?**

2-RSG is the algorithm of Ghadimi and Lan (2013) that takes multiple runs of SGD and picks the best one to get a high probability convergence rate. First, it goes from a mean convergence rate for the iterate with the smallest squared gradient norm to a mean convergence rate for a particular iterate by randomly choosing the iterate. It does this for $\Theta(\log(1/\delta))$ runs of SGD. Then, it uses $\Theta(\log(1/\delta)\sigma^2/\delta\epsilon)$ samples to estimate the $\Theta(\log(1/\delta))$ gradients and pick the run with the smallest squared estimated gradient norm, where σ^2 is the variance of the noise and ϵ is the convergence tolerance. To compete with this, **we introduce a post-processing algorithm that randomly chooses $\Theta(\log(1/\delta))$ iterates of a single run of SGD and then picks the one with the smallest squared estimated gradient norm using $\Theta(\log(1/\delta)\sigma^2/\delta\epsilon)$ samples and we prove that our high probability convergence rate applies to the squared gradient norm of this iterate.** The algorithm and the bound on the squared gradient norm of its output are in Theorem 21. The result can be combined with any high probability convergence rate on a weighted average of the squared gradient norm of the iterates of SGD, such as the results of Li and Orabona (2020), Scaman and Malherbe (2020), and Cutkosky and Mehta (2021). Moreover, the first part of it, going from a high probability bound on a weighted average of random variables to a high probability bound on the smallest of a small subset of those random variables, applies to more general stochastic sequences as shown in Theorem 19, which is a probability result of independent interest.

Underlying our convergence analysis are concentration inequalities. Vladimirova et al. (2020), Wong et al. (2020), and Bakhshizadeh et al. (2023) prove concentration inequalities for sub-Weibull random variables with deterministic scale parameters. However, we need a concentration inequality for a sub-Weibull martingale difference sequence (MDS) where the scale parameters are themselves random variables making up an auxiliary sequence. MDS concentration inequalities upper bound the partial sum of the main sequence and lower bound the partial sum of the auxiliary sequence simultaneously. When the lower bound depends on the partial sum of the main sequence, we call the concentration inequality self-normalized. Freedman (1975) proves a sub-Gaussian MDS concentration inequality. Fan et al. (2015) proves a sub-exponential MDS concentration inequality. Harvey et al. (2019) proves a sub-Gaussian MDS self-normalized concentration inequality. The proofs for these results all rely on the moment generating function (MGF), which does not exist for sub-Weibull random variables with $\theta > 1$, i.e., heavier than sub-exponential. Nevertheless, **we are able to prove a sub-Weibull martingale difference sequence self-normalized concentration inequality.** This is given in Theorem 11. The proof uses the MGF truncation technique of Bakhshizadeh et al. (2023) but applied to an MDS rather than i.i.d. random variables. The truncation level is determined by the tail decay rate and shows up as the $\log(T/\delta)^{\theta-1}$ in the SGD convergence rate.

We also consider convergence under the Polyak-Łojasiewicz (PŁ) inequality. Some examples of problems with PŁ objectives are logistic regression over compact sets (Karimi et al., 2016) and matrix factorization (Sun and Luo, 2016). Deep linear neural networks satisfy the PŁ inequality in large regions of the parameter space (Charles and Papailiopoulos, 2018, Thm. 4.5), as do two-layer neural networks with an extra identity mapping (Li and Yuan, 2017). Furthermore, sufficiently wide neural networks satisfy the PŁ inequality locally around random initialization (Allen-Zhu et al., 2019a,b; Du et al., 2018, 2019; Liu et al., 2022). While strong convexity implies the PŁ inequality, a PŁ function need not even be convex, hence the PŁ condition is considerably more applicable than strong convexity in the context of neural networks. Moreover, as pointed out by Karimi et al. (2016), the convergence proof for gradient descent assuming strong smoothness and strong convexity is actually simplified if we use the PŁ inequality instead of strong convexity. Thus, we would like to find a simple, elegant proof for the high probability convergence of SGD assuming strong smoothness and the PŁ inequality.

Assuming strong smoothness and strong convexity, the tight mean convergence rate of SGD is $O(1/T)$ (Nemirovski et al., 2009). The same mean convergence rate can be shown assuming strong smoothness and the PŁ inequality (Karimi et al., 2016; Orvieto and Lucchi, 2019). But neither of these papers address high probability convergence. To fill this gap, **we prove a $O(\log(1/\delta)/T)$ high probability convergence rate for the objective value of SGD assuming strong smoothness, the PŁ inequality, and norm sub-Gaussian noise.** This is given in Theorem 8. The proof relies on a novel probability result, Theorem 9, that essentially shows that adding two kinds of noise to a contracting sequence—sub-Gaussian noise with variance depending on the sequence itself and sub-exponential noise—results in a sub-exponential sequence. Unfortunately, the proof does not extend to the case of norm sub-Weibull noise. It is unclear if this is just an artifact of the proof or if the faster convergence of SGD under the PŁ inequality really cannot be maintained under norm sub-Weibull noise.

Applicability of assumptions. Given the motivation from neural networks, it is necessary to say a few words regarding the applicability of the Lipschitz continuity and strong smoothness assumptions. First, we show in Lemmas 1 and 2 that these assumptions are satisfied by simple neural network models that are, nevertheless, powerful enough to interpolate generic data sets using only twice the minimal number of parameters required for *any* model to do so. Second, they are satisfied if the iterates remain in a bounded set, which is precisely what happens in the neural tangent kernel setting. In this setting, the local PŁ constant of the least squares loss is sufficiently high (due to overparameterization) for the iterates of SGD to converge to a point close to initialization (Oymak and Soltanolkotabi, 2020). Third, while it is true that Lipschitz continuity and strong smoothness may not be satisfied in practice, our paper is guided by the principle of "how far can we relax assumptions while still being able to prove high probability convergence rates?" An alternative principle is "what can we prove given assumptions that are completely justified in practice?" This is the principle guiding Patel et al. (2022). They provide a neural network counterexample that does not even satisfy (L_0, L_1)-smoothness (Zhang et al., 2020), an alternative assumption to strong smoothness. Then, they prove—assuming only (1) a lower bound on the objective function, (2) local α-Hölder continuity of the gradient, (3) equality of the true and mean gradients, and (4) that the pth moment of the norm of the stochastic gradient, for some $p \in (1, 2]$, is bounded by an upper semi-continuous function—that, with probability 1, either (1) the norm of the iterates of SGD go to infinity, or (2) SGD almost surely converges to a first order stationary point. These assumptions are much weaker than ours, but so is the conclusion. We leave it as a future research direction to further relax our assumptions.

Organization. Section 2 establishes the optimization framework and reviews optimization, probability, and sub-Weibull terminology and results. Section 3 proves the PŁ convergence rate. Section 4 proves the MDS concentration inequality. Section 5 proves the non-convex convergence rate. Section 6 establishes the post-processing algorithm. Section 7 trains a two layer neural network model on synthetic data with Weibull noise injected into the gradient, showing the dependence of the tail of the convergence error on the tail of the noise.

Notation. We use x, y, z for vectors and X, Y, Z, ξ, ψ for random variables. We use e_t for the error vector and distinguish it from Euler's number with the subscript (the subscript will denote the iteration index). We use σ^2 for the variance and so spell out sigma for sigma algebra. We use O and Θ for big-O notation. When comparing two sequences of real numbers, (a_t) and (b_t): $a_t = o(b_t)$ if $\lim a_t/b_t = 0$, $a_t = O(b_t)$ if $\limsup |a_t|/b_t < \infty$, and $a_t = \Theta(b_t)$ if $a_t = O(b_t)$ and $b_t = O(a_t)$. We use $[n]$ to denote the set $\{1, \ldots, n\}$ and Γ to denote the gamma function.

2. Preliminaries

We are interested in the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function, and the SGD iteration

$$x_{t+1} = x_t - \eta_t g_t \ \forall t \in \mathbb{N} \cup \{0\},$$

where $g_t \in \mathbb{R}^d$ is an estimate of $\nabla f(x_t)$, η_t is the step-size, and $x_0 \in \mathbb{R}^d$ is the initial point. We restrict our attention to the setting where x_0 is deterministic, but the results easily extend to the setting where x_0 is a random vector. We define the noise as the vector $e_t = \nabla f(x_t) - g_t$ and assume that, conditioned on the previous iterates, it is unbiased and its norm is sub-Weibull (which, as we will see in Lemma 6, implies that the variance of the noise is bounded).

One of the main examples of Eq. (1) is the stochastic approximation problem: $f = \mathbb{E}[F(\cdot, \xi)]$ where $(\Omega, \mathcal{F}, P)$ is a probability space and $F : \mathbb{R}^d \times \Omega \to \mathbb{R}$ (Nemirovski et al., 2009; Ghadimi and Lan, 2013; Bottou and Bousquet, 2008). In this case, we independently sample $\xi_0, \xi_1, \ldots$ and set $g_t = \nabla F(x_t, \xi_t)$.

Another example is the sample average approximation problem which is the special case of the stochastic approximation problem where there is a finite set $\{\xi_1, \ldots, \xi_n\}$ such that $\xi = \xi_i$ with probability $1/n$. In this setting, we define $F_i = F(\cdot, \xi_i)$ so that

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} F_i(x).$$

2.1 Optimization

We highlight a few key facts we use; see, e.g., Nesterov (2018), for more details. All definitions are with respect to the Euclidean norm $\|\cdot\|$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, and assume $\operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$ is non-empty and denote its minimum by $f^\star$.

If f is continuously differentiable, then f is ρ-Lipschitz if and only if $\|\nabla f(x)\| \leq \rho$ for all $x \in \mathbb{R}^d$. We say f is L-strongly smooth (or L-smooth for short) if its gradient is L-Lipschitz continuous. If f is L-smooth, then a standard result using the Taylor expansion is

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2 \ \forall x, y \in \mathbb{R}^d.$$

Applying this result with $y = x - \frac{1}{L} \nabla f(x)$ and using $f(y) \geq f^\star$, we get

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^\star), \tag{2}$$

or taking $x \in \operatorname{argmin}_{x'} f(x')$, we get

$$f(y) - f^\star \leq \frac{L}{2} \|y - x^\star\|^2.$$

Thus a convergence rate in terms of the iterates is stronger than one in terms of the objective, which is stronger than one in terms of the norm of the gradient. We say f is μ-Polyak-Łojasiewicz (or μ-PŁ for short) if $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^\star)$ $\forall x \in \mathbb{R}^d$. Combining this with Eq. (2) shows that $L \geq \mu$.

To justify the Lipschitz continuity and strong smoothness assumptions in the context of neural networks, we include the following lemmas, proved in Appendix A, which show that both properties are satisfied by the least squares loss applied to the hidden layer of a two layer neural network with, e.g., sigmoid functions such as tanh, arctan, and the logistic function (applying the first lemma) or GELU activation (applying the second lemma). Note that such a model (with fixed outer layer), while a simple example of a neural network,

is already able to interpolate generic data sets with only twice the necessary number of parameters for any model, including deeper ones, to do so (Madden and Thrampoulidis, 2024).

Lemma 1 *Let $m, n, d \in \mathbb{N}$ and $a \in \mathbb{R}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be twice differentiable and assume $|\phi(x)|, |\phi'(x)|, |\phi''(x)| \leq a \ \forall x \in \mathbb{R}$. Let $X \in \mathbb{R}^{d \times n}$, $v \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$. Define*

$$f : \mathbb{R}^{d \times m} \to \mathbb{R} : W \mapsto \frac{1}{2}\|\phi(X^\top W)v - y\|^2.$$

Then f is Lipschitz continuous and strongly smooth.

Lemma 2 *Let $m, n, d \in \mathbb{N}$ and $a, b \in \mathbb{R}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be twice differentiable. Assume $|\phi'(x)| \leq a$ and $|\phi''(x)| \leq b$ for all $x \in \mathbb{R}$. Let $X \in \mathbb{R}^{d \times n}$, $v \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$. Define*

$$f : \mathbb{R}^{d \times m} \to \mathbb{R} : W \mapsto \frac{1}{2}\|\phi(X^\top W)v - y\|^2.$$

Let $\alpha \geq 1$ and define the sublevel set $\mathcal{W} = \{W \in \mathbb{R}^{d \times m} \mid f(W) \leq \alpha\}$. Then, on $\mathcal{W}$, f is

$$a\|v\|_\infty \|X\|_2 \sqrt{2\alpha m}\text{-Lipschitz continuous and}$$
$$\left(a^2\|v\|_\infty^2 \|X\|_2^2 m + b\|v\|_\infty \|X\|_2 \|X\|_{1,2}\sqrt{2\alpha}\right)\text{-strongly smooth.}$$

2.2 Probability

See Section 20 of Billingsley (1995) for details on the convergence of random variables. A sequence of random variables (X_t) converges to a random variable X:

- in probability if, for all $\epsilon > 0$, $\lim_t P(|X_t - X| > \epsilon) = 0$, denoted $X_t \xrightarrow{p} X$;
- in mean if $\lim_t \mathbb{E}|X_t - X| = 0$, denoted $X_t \xrightarrow{L^1} X$;
- almost surely if $P(\lim_t X_t = X) = 1$, denoted $X_t \xrightarrow{a.s.} X$.

When the rate of convergence is of interest, we say a sequence of random variables (X_t) converges to X:

- with mean convergence rate $r(t)$ if, for all t, $\mathbb{E}|X_t - X| \leq r(t)$;
- with high probability convergence rate $\tilde{r}(t, \delta)$ if, for all t and δ, $P\left(|X_t - X| \leq \tilde{r}(t, \delta)\right) \geq 1 - \delta$.

All five kinds of convergence are interrelated:

- Convergence in mean and convergence almost surely both imply convergence in probability.
- Mean convergence with rate $r(t)$ such that $r(t) \to 0$ as $t \to \infty$ implies convergence in mean.

- By Markov's inequality, mean convergence with rate $r(t)$ implies high probability convergence with rate $\tilde{r}(t, \delta) = r(t)/\delta$.

- By the Borel-Cantelli lemma (Billingsley, 1995, Thm. 4.3), high probability convergence with rate $\tilde{r}(t, \delta) = r(t)p(\delta)$ implies almost sure convergence if, for all $a > 0$, $\sum_{t=0}^{\infty} \min\{1, p^{-1}(a/r(t))\} < \infty$. If $p(\delta) = 1/\delta^c$ for some $c > 0$, then $r(t) = o(1/t^c)$ is required. If $p(\delta) = \log(1/\delta)$, then only $r(t) = O(1/t^c)$ for some $c > 0$ is required.

See Section 35 of Billingsley (1995) for details on martingale difference sequences. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A sequence, $(\mathcal{F}_i)$, of nested sigma algebras in $\mathcal{F}$ (i.e., $\mathcal{F}_i \subset \mathcal{F}_{i+1} \subset \mathcal{F}$) is called a filtration, in which case $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ is called a filtered probability space. A sequence of random variables (ξ_i) is said to be adapted to $(\mathcal{F}_i)$ if each ξ_i is $\mathcal{F}_i$-measurable. Furthermore, if $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = \xi_{i-1}\ \forall i$, then (ξ_i) is called a martingale. On the other hand, if $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = 0\ \forall i$, then (ξ_i) is called a martingale difference sequence.

For the noise sequence (e_t), we define a corresponding filtration $(\mathcal{F}_t)$ by letting $\mathcal{F}_t$ be the sigma algebra generated by $e_0, \ldots, e_t$ for all $t \geq 0$ and setting $\mathcal{F}_{-1} = \{\emptyset, \Omega\}$. Note that x_t is $\mathcal{F}_{t-1}$-measurable while e_t is $\mathcal{F}_t$-measurable.

2.3 Sub-Weibull Random Variables

We consider sub-Gaussian, sub-exponential, and sub-Weibull random variables.

Definition 3 *A random variable X is K-sub-Gaussian if $\mathbb{E}\left[\exp\left(X^2/K^2\right)\right] \leq 2$.*

Definition 4 *A random variable X is K-sub-exponential if $\mathbb{E}\left[\exp\left(|X|/K\right)\right] \leq 2$.*

Definition 5 *A random variable X is K-sub-Weibull(θ) if $\mathbb{E}\left[\exp\left((|X|/K)^{1/\theta}\right)\right] \leq 2$.*

See Proposition 2.5.2 of Vershynin (2018) for equivalent definitions of sub-Gaussian, Proposition 2.7.1 of Vershynin (2018) for equivalent definitions of sub-exponential, and Theorem 2.1 of Vladimirova et al. (2020) for equivalent definitions of sub-Weibull. Note that sub-Gaussian and sub-exponential are special cases of sub-Weibull using $\theta = \frac{1}{2}$ and $\theta = 1$, respectively. The tail parameter θ measures the heaviness of the tail—higher values correspond to heavier tails—and the scale parameter K gives us the following bound on the second moment.

Lemma 6 *If X is K-sub-Weibull(θ) then $\mathbb{E}\left[|X|^p\right] \leq 2\Gamma(\theta p + 1)K^p\ \forall p > 0$. In particular, $\mathbb{E}\left[X^2\right] \leq 2\Gamma(2\theta + 1)K^2$.*

Proof First, for all $t \geq 0$,

$$\begin{aligned} P\left(|X| \geq t\right) &= P\left(\exp\left((|X|/K)^{1/\theta}\right) \geq \exp\left((t/K)^{1/\theta}\right)\right) \\ &\leq 2\exp\left(-\left(t/K\right)^{1/\theta}\right). \end{aligned}$$

Second,

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty P\left(|X|^p \geq x\right) dx \\
&\leq 2\int_0^\infty \exp\left(-\left(x^{1/p}/K\right)^{1/\theta}\right) \\
&= 2\theta p K^p \int_0^\infty \exp(-u) u^{\theta p - 1} du \\
&= 2\theta p \Gamma(\theta p) K^p \\
&= 2\Gamma(\theta p + 1) K^p.
\end{aligned}$$

■

The proof demonstrates the general techniques for going between probability and expectation in this type of analysis. To go from probability to expectation, the CDF formula is used:

$$\mathbb{E}\left[|Y|\right] = \int_0^\infty P\left(|Y| > t\right) dt.$$

To go from expectation to probability, Markov's inequality is used:

$$P\left(|Y| > t\right) \leq \frac{1}{t}\mathbb{E}\left[|Y|\right] \ \ \forall t > 0.$$

In both cases, the trick is choosing what Y should be, e.g. $Y = |X|^p$ or $Y = \exp((\lambda|X|)^{1/\theta})$.

Remark 7 *Note that our definitions of sub-Gaussian, sub-exponential, and sub-Weibull do not require the random variable to be centered. Thus, we do not center the constant random variable $X \equiv x$ to think of it as having scale parameter zero. Instead, for each $\theta > 0$, $X \equiv x$ is $|x|/\log(2)^\theta$-sub-Weibull(θ).*

We can apply Remark 7 in the following way to show how the sub-Weibull scale parameter might scale with the tail parameter for the gradient noise in the sample average approximation setting where $f(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$. Assume that $\|\nabla f(x) - \nabla F_i(x)\| \leq \rho$ for all $x \in \mathcal{X} \subset \mathbb{R}^d$ and all $i \in [n]$. Since $[n]$ is a finite set, this follows for some $\rho > 0$ if $\mathcal{X}$ is compact. Then we get that $\|\nabla f(x) - \nabla F_\xi(x)\|$ is $\rho/\log(2)^\theta$-sub-Weibull(θ) for all $x \in \mathcal{X}$ and all $\theta > 0$. Thus, we can decrease the scale parameter by increasing the tail parameter. However, the convergence rate in Theorem 15 has a coefficient of $\theta^{2\theta}K^2$ when $\theta > 1/2$, and $\theta^{2\theta}\rho^2/\log(2)^{2\theta}$ increases as θ increases. So, we cannot hope to erase the noise in the convergence rate by increasing the scale parameter. On the other hand, it may be the case that there is some θ slightly larger than $1/2$ which minimizes the convergence bound when all of the constants are considered.

3. PŁ Convergence

In this section, we prove the following convergence bound, proving Theorem 9, a probability result, along the way. The convergence bound matches the optimal convergence rate in mean of $O(1/T)$ and has $\log(1/\delta)$ dependence on δ.

Theorem 8 *Assume f is L-smooth and μ-PŁ and that, conditioned on the previous iterates, e_t is centered and $\|e_t\|$ is σ-sub-Gaussian. Then, SGD with step-size*

$$\eta_t = \frac{2t + 4\kappa - 1}{\mu(t + 2\kappa)^2},$$

where $\kappa = L/\mu$, constructs a sequence (x_t) such that, w.p. $\geq 1 - \delta$ for all $\delta \in (0, 1)$,

$$f(x_T) - f^\star = O\left(\frac{L\sigma^2 \log(e/\delta)}{\mu^2 T}\right).$$

Proof Assume f is L-smooth, hence

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 \tag{3}$$

for all $x, y \in \mathbb{R}^d$. Set $y = x_{t+1} = x_t - \eta_t g_t = x_t - \eta_t(\nabla f(x_t) - e_t)$ and $x = x_t$ and subtract $f^\star$. Additionally assume f is μ-PŁ and $\eta_t \leq 1/L$. Then

$$\begin{aligned}
&f(x_{t+1}) - f^\star \\
&\quad\leq f(x_t) - f^\star - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla f(x_t)\|^2 + \eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle + \frac{L\eta_t^2}{2}\|e_t\|^2 \\
&\quad\leq f(x_t) - f^\star - \frac{\eta_t}{2}\|\nabla f(x_t)\|^2 + \eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle + \frac{L\eta_t^2}{2}\|e_t\|^2 \\
&\quad\leq (1 - \mu\eta_t)(f(x_t) - f^\star) + \eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle + \frac{L\eta_t^2}{2}\|e_t\|^2
\end{aligned} \tag{4}$$

where the second inequality used $\eta_t \leq 1/L$ and the third used the PŁ inequality.

Note that

$$\eta_t = \frac{2t + 4\kappa - 1}{\mu(t + 2\kappa)^2} = \Theta\left(\frac{1}{\mu t}\right)$$

and, since η_t is decreasing in t for $t \geq 0$,

$$\eta_t \leq \frac{4\kappa - 1}{\mu(2\kappa)^2} \leq \frac{4\kappa}{\mu 4\kappa^2} = \frac{1}{L},$$

so we can apply Eq. (4).

Define

$$\begin{aligned}
X_t &= (t + 2\kappa - 1)^2(f(x_t) - f^\star) \\
Y_t &= \eta_t(1 - L\eta_t)(t + 2\kappa)^2\langle \nabla f(x_t), e_t\rangle \\
Z_t &= \frac{L\eta_t^2(t + 2\kappa)^2}{2}\|e_t\|^2.
\end{aligned}$$

Then multiplying both sides of Eq. (4) by $(t + 2\kappa)^2$ and noting $1 - \mu\eta_t = \frac{(t+2\kappa-1)^2}{(t+2\kappa)^2}$ we get the recursion

$$X_{t+1} \leq X_t + Y_t + Z_t.$$

Recall that we defined $\mathcal{F}_t$ as the sigma algebra generated by $e_0, \ldots, e_t$ for all $t \geq 0$ and $\mathcal{F}_{-1} = \{\emptyset, \Omega\}$. So, X_t is $\mathcal{F}_{t-1}$-measurable, Y_t is $\mathcal{F}_t$-measurable, and Z_t is $\mathcal{F}_t$-measurable.

For all $t \geq 0$, assume, conditioned on $\mathcal{F}_{t-1}$, that e_t is centered and $\|e_t\|$ is σ-sub-Gaussian. Then, Y_t is centered conditioned on $\mathcal{F}_{t-1}$. Note that we can bound Y_t using Cauchy-Schwarz and Eq. (2) in the following way:

$$\langle \nabla f(x_t), e_t \rangle \leq \|\nabla f(x_t)\| \cdot \|e_t\| \leq \sqrt{2L(f(x_t) - f^\star)}\|e_t\|.$$

But, if we use this to get a new recursion, then we get a sub-optimal convergence rate. Instead, we need to keep the same recursion but use the bound on Y_t in its MGF:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{Y_t^2}{\frac{18L}{\mu^2}X_t\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[\exp\left(\frac{\eta_t^2(1-L\eta_t)^2(t+2\kappa)^4\|\nabla f(x_t)\|^2\|e_t\|^2}{\frac{18L}{\mu^2}X_t\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\eta_t^2(1-L\eta_t)^2\frac{(t+2\kappa)^4}{(t+2\kappa-1)^2}2LX_t\|e_t\|^2}{\frac{18L}{\mu^2}X_t\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \text{ via (2)} \\
&\leq \mathbb{E}\left[\exp\left(\frac{\frac{18L}{\mu^2}X_t\|e_t\|^2}{\frac{18L}{\mu^2}X_t\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leq 2.
\end{aligned}$$

Thus, Y_t is $\frac{18L}{\mu^2}X_t\sigma^2$-sub-Gaussian conditioned on $\mathcal{F}_{t-1}$. Similarly,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{Z_t}{\frac{2L}{\mu^2}\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[\exp\left(\frac{\frac{2L}{\mu^2}\|e_t\|^2}{\frac{2L}{\mu^2}\sigma^2}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leq 2.
\end{aligned}$$

and so Z_t is $\frac{2L}{\mu^2}\sigma^2$-sub-exponential conditioned on $\mathcal{F}_{t-1}$.

So, we have a recursion for a stochastic sequence with two types of additive noise—sub-Gaussian noise, with variance depending on the sequence itself, and sub-exponential noise. We prove that this makes the main sequence sub-exponential in the following theorem.

Theorem 9 *Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (X_{i+1}), (Y_i), and (Z_i) be adapted to $(\mathcal{F}_i)$, and X_0 be deterministic. Let (α_i) and (β_i) be sequences of non-negative reals and let (γ_i) be a sequence of positive reals. Assume X_i and Z_i are non-negative almost surely. Assume $\mathbb{E}[\exp(\lambda Y_i) \mid \mathcal{F}_{i-1}] \leq \exp(\frac{\lambda^2}{2}\beta_i^2 X_i)$ for all $\lambda \in \mathbb{R}$ and $\mathbb{E}[\exp(\lambda Z_i) \mid \mathcal{F}_{i-1}] \leq \exp(\lambda\gamma_i)$ for all $\lambda \in [0, \gamma_i^{-1}]$. Assume*

$$X_{i+1} \leq \alpha_i X_i + Y_i + Z_i.$$

Then, for any sequence of positive reals (K_i) such that $K_0 \geq X_0$ and, for all $i \geq 0$, $K_{i+1}^2 \geq (\alpha_i K_i + 2\gamma_i)K_{i+1} + \beta_i^2 K_i$, and for any $n \geq 0$, we have, w.p. $\geq 1 - \delta$ for all $\delta \in (0, 1)$,

$$X_n \leq K_n \log(e/\delta).$$

Proof We want to find requirements on a sequence (K_i) such that $\mathbb{E}\left[\exp(\lambda X_i)\right] \leq \exp(\lambda K_i)\ \forall \lambda \in [0, K_i^{-1}]$. Then, by Markov's inequality and taking $\lambda = K_n^{-1}$, $P\left(X_n \geq \log(e/\delta)K_n\right) = P\left(\exp(X_n/K_n) \geq e/\delta\right) \leq \delta$. Our proof is inductive. For the base case, we need

$$(1)\ K_0 \geq X_0.$$

For the induction step, assume $\mathbb{E}\left[\exp(\tilde{\lambda}X_i)\right] \leq \exp(\tilde{\lambda}K_i)\ \forall \tilde{\lambda} \in [0, K_i^{-1}]$. Let $\lambda \in [0, K_{i+1}^{-1}]$. Then

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda X_{i+1})\right] &\leq \mathbb{E}\left[\exp(\lambda\alpha_i X_i + \lambda Y_i + \lambda Z_i)\right] \quad \text{by non-negativity} \\
&\overset{\text{TE}}{=} \mathbb{E}\left[\exp(\lambda\alpha_i X_i)\mathbb{E}\left[\exp(\lambda Y_i)\exp(\lambda Z_i) \mid \mathcal{F}_{i-1}\right]\right] \\
&\overset{\text{CS}}{\leq} \mathbb{E}\left[\exp(\lambda\alpha_i X_i)\mathbb{E}\left[\exp(2\lambda Y_i) \mid \mathcal{F}_{i-1}\right]^{1/2}\mathbb{E}\left[\exp(2\lambda Z_i) \mid \mathcal{F}_{i-1}\right]^{1/2}\right] \\
&\overset{(2)}{\leq} \mathbb{E}\left[\exp(\lambda\alpha_i X_i)\exp\left(\frac{(2\lambda)^2}{2}\beta_i^2 X_i\right)^{1/2}\exp(2\lambda\gamma_i)^{1/2}\right] \quad \text{if} \quad 2\lambda \in [0, \gamma_i^{-1}] \\
&= \mathbb{E}\left[\exp(\lambda\alpha_i X_i + \lambda^2\beta_i^2 X_i + \lambda\gamma_i)\right] \\
&= \mathbb{E}\left[\exp(\lambda(\alpha_i + \lambda\beta_i^2)X_i\right]\exp(\lambda\gamma_i) \\
&\overset{(3)}{\leq} \exp(\lambda((\alpha_i + \lambda\beta_i^2)K_i + \gamma_i)) \quad \text{via induction if } \tilde{\lambda} := \lambda(\alpha_i + \lambda\beta_i^2) \leq K_i^{-1} \\
&\overset{(4)}{\leq} \exp(\lambda K_{i+1})
\end{aligned}$$

where TE denotes the law of total expectation, CS denotes the Cauchy-Schwarz inequality, and $(2) - (4)$ are the requirements

$$\begin{aligned}
&(2)\ 2\lambda \leq \gamma_i^{-1} \impliedby 2K_{i+1}^{-1} \leq \gamma_i^{-1} \iff K_{i+1} \geq 2\gamma_i \\
&(3)\ \lambda(\alpha_i + \lambda\beta_i^2) \leq K_i^{-1} \impliedby K_{i+1}^{-1}(\alpha_i + K_{i+1}^{-1}\beta_i^2) \leq K_i^{-1} \iff K_{i+1}^2 \geq \alpha_i K_i K_{i+1} + \beta_i^2 K_i \\
&(4)\ K_{i+1} \geq (\alpha_i + \lambda\beta_i^2)K_i + \gamma_i \impliedby K_{i+1} \geq (\alpha_i + K_{i+1}^{-1}\beta_i^2)K_i + \gamma_i \\
&\qquad\qquad \iff K_{i+1}^2 \geq (\alpha_i K_i + \gamma_i)K_{i+1} + \beta_i^2 K_i.
\end{aligned}$$

The assumptions of the theorem imply requirements $(2) - (4)$, completing the proof. ■

The proof is similar to the proof of Theorem 4.1 of Harvey et al. (2019) but with some key differences. There, the recursion is $X_{i+1} \leq \alpha_i X_i + Y_i\sqrt{X_i} + \gamma_i$ where $\mathbb{E}\left[\exp(\lambda Y_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\beta_i^2\right)$ and γ_i is deterministic. We had to move the implicit dependence of the sub-Gaussian term inside of the MGF. We also had to allow γ_i to be a sub-exponential random variable Z_i, and so applied Cauchy-Schwarz, which contributed the 2 in the recursion for (K_i).

Note that if $\alpha_i = 1, \gamma_i = \gamma, \beta_i = \beta\ \forall i$, then the assumptions on (K_i) are satisfied by, for example, $K_0 = X_0$ and $K_{i+1} = K_i + 2\gamma + \beta^2\ \forall i$. Returning to the proof of Theorem 8, for absolute constants a and b, by Proposition 2.5.2 and Proposition 2.7.1 of Vershynin (2018),

we can set $\beta_t^2 = \frac{18aL\sigma^2}{\mu^2}$ and $\gamma_t = \frac{2bL\sigma^2}{\mu^2}$ and apply Theorem 9 with

$$\begin{aligned} K_T &:= X_0 + \sum_{t=0}^{T-1} (2\gamma_t + \beta_t^2) \\ &= X_0 + \frac{(18a + 4b)L\sigma^2 T}{\mu^2}. \end{aligned}$$

Dividing by $(T + 2\kappa)^2$ completes the proof. ■

Remark 10 *It is natural to ask whether we can relax the sub-Gaussian assumption to a sub-Weibull assumption. In Theorem 9, we need a bound on the MGFs of both Y_i and Z_i. But, if $\|e_t\|$ is σ-sub-Weibull(θ) with $\theta > 1/2$, then $\|e_t\|^2$ is σ^2-sub-Weibull(2θ). Thus, Z_i would be sub-Weibull with tail parameter greater than 1, and so may have an infinite moment generating function for all $\lambda > 0$.*

A big take-away from the PŁ analysis is its simplicity, matching the simplicity of gradient descent's convergence analysis in the same setting. It also serves as a good warm-up for the non-convex analysis that follows. Induction on the convergence recursion does not work for the non-convex analysis. Instead it requires an MDS self-normalized concentration inequality, which we state and prove in the next section.

4. Concentration Inequality

In this section, we prove the following MDS concentration inequality. For $\theta = 1/2$ without α (i.e., $\alpha = 0$), Eq. (6), we recover the classical Freedman's inequality (Freedman, 1975). For $\theta \in (1/2, 1]$ without α, Eq. (8), we recover Theorem 2.6 of Fan et al. (2015). For $\theta = 1/2$ with α, Eq. (5), we recover Theorem 3.3 of Harvey et al. (2019), called the "Generalized Freedman" inequality.

Theorem 11 *Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$. For all $i \in [n]$, assume $K_{i-1} \geq 0$ almost surely, $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = 0$, and*

$$\mathbb{E}\left[\exp\left((|\xi_i|/K_{i-1})^{1/\theta}\right) \mid \mathcal{F}_{i-1}\right] \leq 2$$

where $\theta \geq 1/2$. If $\theta > 1/2$, assume there exist constants (m_i) such that $K_{i-1} \leq m_i$ almost surely for all $i \in [n]$.

If $\theta = 1/2$, then for all $x, \beta \geq 0$, and $\alpha > 0$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} 2K_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta \right\}\right) \leq \exp(-\lambda x + 2\lambda^2 \beta), \tag{5}$$

and for all $x, \beta, \lambda \geq 0$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} 2K_{i-1}^2 \leq \beta \right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right). \tag{6}$$

If $\theta \in \left(\frac{1}{2}, 1\right]$, *define*

$$a = (4\theta)^{2\theta} e^2$$
$$b = (4\theta)^{\theta} e.$$

For all $x, \beta \geq 0$, *and* $\alpha \geq b \max_{i\in[n]} m_i$, *and* $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta \right\}\right) \leq \exp(-\lambda x + 2\lambda^2 \beta), \tag{7}$$

and for all $x, \beta \geq 0$, *and* $\lambda \in \left[0, \frac{1}{b \max_{i\in[n]} m_i}\right]$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta \right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right). \tag{8}$$

If $\theta > 1$, *let* $\delta \in (0, 1)$. *Define*

$$a = (2^{2\theta+1} + 2)\Gamma(2\theta + 1) + \frac{2^{3\theta}\Gamma(3\theta + 1)}{3\log(n/\delta)^{\theta-1}}$$
$$b = 2\log(n/\delta)^{\theta-1}.$$

For all $x, \beta \geq 0$, *and* $\alpha \geq b \max_{i\in[n]} m_i$, *and* $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta \right\}\right) \leq \exp(-\lambda x + 2\lambda^2 \beta) + 2\delta,$$

and for all $x, \beta \geq 0$, *and* $\lambda \in \left[0, \frac{1}{b \max_{i\in[n]} m_i}\right]$,

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta \right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right) + 2\delta. \tag{9}$$

Proof Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$. For all $i \in [n]$, assume $0 \leq K_{i-1} \leq m_i$ almost surely, $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = 0$, and

$$\mathbb{E}\left[\exp\left((|\xi_i|/K_{i-1})^{1/\theta}\right) \mid \mathcal{F}_{i-1}\right] \leq 2.$$

What we want to upper bound in this setting is

$$P\left(\bigcup_{k\in[n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} K_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta \right\}\right)$$

for constants $x, \alpha > 0$ and $\beta \geq 0$. To understand how to use this, set $\beta = 0$ and observe

$$\begin{aligned}
&P\left(\sum_{i=1}^{n} \xi_i \geq x + \frac{1}{\alpha}\sum_{i=1}^{n} K_{i-1}^2\right) \\
&\leq P\left(\sum_{i=1}^{n} \xi_i \geq x \text{ and } \sum_{i=1}^{n} K_{i-1}^2 \leq \alpha \sum_{i=1}^{n} \xi_i\right) \\
&\leq P\left(\bigcup_{k\in[n]} \left\{\sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} K_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i\right\}\right)
\end{aligned}$$

by monotonicity ($A \subset B \implies P(A) \leq P(B)$). The stronger bound over the union of partial sum bounds does not help us since the constants cannot depend on k. The union of partial sum bounds arises in the proof from a stopping time defined as the first k such that the corresponding set occurs.

Proving such a bound would typically involve using an MGF bound for ξ_i, but the MGF is infinite in our setting. To get around this, we truncate ξ_i to an appropriate level. By accounting for the probability of exceeding truncation, applying an MGF bound for truncated random variables, and applying a concentration inequality for bounded MGF martingale different sequence, we are able to prove a concentration inequality for sub-Weibull martingale difference sequences.

Probability of exceeding truncation. Let $\delta \in (0, 1)$. We want to define the truncation level so that the probability of any of the ξ_i exceeding it is smaller than $O(\delta)$. This ends up being $\widetilde{\xi}_i = \xi_i \mathbb{I}_{\{\xi_i \leq cK_{i-1}\}}$ with $c = \log(n/\delta)^\theta$ as we will see.

First, for any $c > 0$,

$$\begin{aligned}
P\left(|\xi_i| \geq cK_{i-1}\right) &= P\left(\exp\left(\left(|\xi_i|/K_{i-1}\right)^{1/\theta}\right) \geq \exp\left(c^{1/\theta}\right)\right) \\
&\leq \exp\left(-c^{1/\theta}\right) \mathbb{E}\left[\exp\left(\left(|\xi_i|/K_{i-1}\right)^{1/\theta}\right)\right] \\
&= \exp\left(-c^{1/\theta}\right) \mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(|\xi_i|/K_{i-1}\right)^{1/\theta}\right) \mid \mathcal{F}_{i-1}\right]\right] \\
&\leq 2\exp\left(-c^{1/\theta}\right)
\end{aligned}$$

where we use Markov's inequality, the law of total expectation, and the sub-Weibull assumption. In particular, $P\left(\xi_i > \log(n/\delta)^\theta K_{i-1}\right) \leq 2\delta/n$.

Then, we can bound

$$P\left(\bigcup_{k\in[n]} \left\{\sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} K_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta\right\}\right)$$

by

$$P\left(\bigcup_{k\in[n]} \left\{\sum_{i=1}^{k} \widetilde{\xi}_i \geq x \text{ and } \sum_{i=1}^{k} K_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \widetilde{\xi}_i + \beta\right\}\right)$$

and

$$P\left(\bigcup_{i\in[n]}\left\{\xi_i > \log(n/\delta)^\theta K_{i-1}\right\}\right) \le \sum_{i=1}^{n} P\left(\xi_i > \log(n/\delta)^\theta K_{i-1}\right) \le 2\delta.$$

and so proceed using $\widetilde{\xi}_i$ while carrying around an additional 2δ probability.

MGF bound of truncated random variable. In order to bound the MGF of our truncated random variables, we prove the following lemma which slightly modifies Corollary 2 of Bakhshizadeh et al. (2023) using the law of total expectation. The main subtlety in the extension is where we use the following bound

$$\begin{aligned} P\left(|X| \ge t \mid \mathcal{G}\right) &= P\left(\exp\left((|X|/K_0)^{1/\theta}\right) \ge \exp\left((t/K_0)^{1/\theta}\right) \mid \mathcal{G}\right) \\ &\le \exp\left(-(t/K_0)^{1/\theta}\right) \mathbb{E}\left[\exp\left((|X|/K_0)^{1/\theta}\right) \mid \mathcal{G}\right] \\ &\le 2\exp\left(-(t/K_0)^{1/\theta}\right) \end{aligned}$$

which we denote by $*$. Otherwise, the proof exactly follows theirs.

Lemma 12 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G} \subset \mathcal{F}$ be a sigma algebra, and X and K_0 be random variables. Assume K_0 is $\mathcal{G}$-measurable. Assume, conditioned on $\mathcal{G}$, that X is centered and K_0-sub-Weibull(θ) with $\theta > 1$. Define $\widetilde{X} = X\mathbb{I}_{\{X \le cK_0\}}$. Then*

$$\mathbb{E}\left[\exp(\lambda\widetilde{X}) \mid \mathcal{G}\right] \le \exp\left(\frac{\lambda^2}{2} a K_0^2\right) \ \forall \lambda \in \left[0, \frac{1}{bK_0}\right]$$

where

$$\begin{aligned} a &= (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3c^{1-1/\theta}} \\ b &= 2c^{1-1/\theta}. \end{aligned}$$

Proof For convenience, define $L_0 = cK_0$. Let $\lambda \in \left[0, \frac{1}{bK_0}\right]$. That is, $\lambda \ge 0$ and $\lambda L_0^{1-1/\theta} K_0^{1/\theta} \le \frac{1}{2}$. Since $\lambda \ge 0$, we have, by Lemma 4 of Bakhshizadeh et al. (2023),

$$\mathbb{E}\left[\exp(\lambda\widetilde{X}) \mid \mathcal{G}\right] \le \exp\left(\frac{\lambda^2}{2}\left(\mathbb{E}\left[X^2\mathbb{I}_{\{X<0\}} \mid \mathcal{G}\right] + \mathbb{E}\left[X^2\exp(\lambda X)\mathbb{I}_{\{0\le X\le L_0\}} \mid \mathcal{G}\right]\right)\right).$$

Observe

$$\begin{aligned} \mathbb{E}\left[X^2\mathbb{I}_{\{X<0\}} \mid \mathcal{G}\right] &= \int_0^\infty P\left(X^2\mathbb{I}_{\{X<0\}} > x \mid \mathcal{G}\right) dx \\ &= \int_0^\infty P\left(X^2 > t^2, X < 0 \mid \mathcal{G}\right) 2t\, dt \\ &\le \int_0^\infty P\left(|X| > t \mid \mathcal{G}\right) 2t\, dt \\ &\overset{*}{\le} 2\int_0^\infty \exp\left(-(t/K_0)^{1/\theta}\right) 2t\, dt \\ &= 2\Gamma(2\theta+1)K_0^2 \end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}\left[X^2\exp(\lambda X)\mathbb{I}_{\{0\le X\le L_0\}} \mid \mathcal{G}\right]\\
&=\int_0^\infty P\left(X^2\exp(\lambda X)\mathbb{I}_{\{0\le X\le L_0\}} > x \mid \mathcal{G}\right)dx\\
&=\int_0^\infty P\left(X^2\exp(\lambda X) > t^2\exp(\lambda t), 0\le X\le L_0 \mid \mathcal{G}\right)(2t+\lambda t^2)\exp(\lambda t)dt\\
&=\int_0^\infty P\left(|X| > t, 0\le X\le L_0 \mid \mathcal{G}\right)(2t+\lambda t^2)\exp(\lambda t)dt\\
&\le\int_0^{L_0} P\left(|X| > t \mid \mathcal{G}\right)(2t+\lambda t^2)\exp(\lambda t)dt\\
&\overset{*}{\le} 2\int_0^{L_0}\exp\left(-\left(1-\lambda t^{1-1/\theta}K_0^{1/\theta}\right)(t/K_0)^{1/\theta}\right)(2t+\lambda t^2)dt\\
&\le 2\int_0^{L_0}\exp\left(-\left(1-\lambda L_0^{1-1/\theta}K_0^{1/\theta}\right)(t/K_0)^{1/\theta}\right)(2t+\lambda t^2)dt\\
&=2\int_0^{L_0}\exp(-u)\left(\frac{2K_0^2\theta u^{2\theta-1}}{\left(1-\lambda L_0^{1-1/\theta}K_0^{1/\theta}\right)^{2\theta}}+\frac{\lambda K_0^3\theta u^{3\theta-1}}{\left(1-\lambda L_0^{1-1/\theta}K_0^{1/\theta}\right)^{3\theta}}\right)du\\
&=\frac{2K_0^2\Gamma(2\theta+1)}{\left(1-\lambda L_0^{1-1/\theta}K_0^{1/\theta}\right)^{2\theta}}+\frac{2\lambda K_0^3\Gamma(3\theta+1)}{3\left(1-\lambda L_0^{1-1/\theta}K_0^{1/\theta}\right)^{3\theta}}\\
&\le\left(2^{2\theta+1}\Gamma(2\theta+1)+\frac{2^{3\theta}K_0\Gamma(3\theta+1)}{3L_0^{1-1/\theta}K_0^{1/\theta}}\right)K_0^2\\
&=\left(2^{2\theta+1}\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3\log(n/\delta)^{\theta-1}}\right)K_0^2.
\end{aligned}
$$

■

Applying this to our setting, we get the MGF bound

$$\mathbb{E}\left[\exp(\lambda\widetilde{\xi}_i) \mid \mathcal{F}_{t-1}\right]\le\exp\left(\frac{\lambda^2}{2}aK_{i-1}^2\right)\ \forall\lambda\in\left[0,\frac{1}{bK_{i-1}}\right] \tag{10}$$

where

$$
\begin{aligned}
a&=(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3\log(n/\delta)^{\theta-1}}\\
b&=2\log(n/\delta)^{\theta-1}.
\end{aligned}
$$

Concentration inequality for MGF bound. Now we just need a self-normalized concentration inequality in the following setting:

Let $(\Omega,\mathcal{F},(\mathcal{F}_i),P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to (F_i). Let $n\in\mathbb{N}$. For all $i\in[n]$, assume $0\le K_{i-1}\le m_i$ almost surely, $\mathbb{E}\left[\xi \mid \mathcal{F}_{i-1}\right]=0$, and

$$\mathbb{E}\left[\exp(\lambda\xi_i) \mid \mathcal{F}_{i-1}\right]\le\exp\left(\frac{\lambda^2}{2}aK_{i-1}^2\right)\ \forall\lambda\in\left[0,\frac{1}{bK_{i-1}}\right]$$

for constants $a, b > 0$.

We can recognize this as a centered sub-exponential type MGF bound from Proposition 2.7.1 of Vershynin (2018). Theorem 2.6 of Fan et al. (2015) proves a concentration inequality in this setting, but not a self-normalized one. On the other hand, Theorem 3.3 of Harvey et al. (2019) proves a self-normalized concentration inequality in the centered sub-Gaussian setting, thus extending the original result of Freedman (1975) to a self-normalized result. We use the same trick of Harvey et al. (2019) to extend Theorem 2.6 of Fan et al. (2015) to a self-normalized result.

We distill the proof technique of Fan et al. (2015) into the following lemma so that we can apply it in our setting.

Lemma 13 *(Fan et al., 2015, Pf. of Thm. 2.1) Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. If (ψ_i) is adapted to $(\mathcal{F}_i)$ and (A_k) is a sequence of events; then*

$$P\left(\bigcup_{k\in[n]} A_k\right) \leq \sup_{k\in[n]} \mathbb{I}_{A_k} \prod_{i=1}^{k} \frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i}.$$

Proof Defining $Z_k = \prod_{i=1}^{k} \frac{\psi_i}{\mathbb{E}[\psi_i|\mathcal{F}_{i-1}]}$, then (Z_k) is a martingale. Let T be a stopping time. Then the stopped process $(Z_{k\wedge T})$ is a martingale (where $a \wedge b$ denotes $\min\{a, b\}$). Moreover, $Z_{k\wedge T}$ is a probability density so define the conjugate probability measure $dP' = Z_{n\wedge T} dP$.

Define the stopping time $T(\omega) = \min\{k \in [n] \mid \omega \in A_k\}$. Then $\mathbb{I}_{\cup_{k\in[n]} A_k} = \sum_{i=1}^{n} \mathbb{I}_{\{T=k\}}$.

Observe,

$$\begin{aligned}
P\left(\bigcup_{k\in[n]} A_k\right) &= \mathbb{E}'\left[Z_{n\wedge T}^{-1} \sum_{k=1}^{n} \mathbb{I}_{\{T=k\}}\right] \\
&= \sum_{k=1}^{n} \mathbb{E}'\left[\prod_{i=1}^{k} \frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i} \mathbb{I}_{\{T=k\}}\right] \\
&\leq \left(\sup_{k\in[n]} \mathbb{I}_{A_k} \prod_{i=1}^{k} \frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i}\right) \sum_{k=1}^{n} \mathbb{E}'\left[\mathbb{I}_{\{T=k\}}\right] \\
&= \sup_{k\in[n]} \mathbb{I}_{A_k} \prod_{i=1}^{k} \frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i}.
\end{aligned}$$

■

We apply Lemma 13 to our setting to get the following self-normalized concentration inequality.

Lemma 14 *Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to (F_i). Let $n \in \mathbb{N}$. For all $i \in [n]$, assume $0 \leq K_{i-1} \leq m_i$ almost surely, $\mathbb{E}\left[\xi \mid \mathcal{F}_{i-1}\right] = 0$, and*

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} a K_{i-1}^2\right) \ \forall \lambda \in \left[0, \frac{1}{b K_{i-1}}\right]$$

for constants $a, b > 0$. Then, for all $x, \beta \geq 0$, and $\alpha \geq b \max_{i\in[n]} m_i$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta).$$

Proof By Claim C.2 of Harvey et al. (2019), if $0 \leq \lambda \leq \frac{1}{2\alpha}$, then $\exists\ c \in [0,2]$ such that $\frac{1}{2}(\lambda + \alpha c\lambda^2)^2 = c\lambda^2$. Define

$$\psi_i = \exp\left((\lambda + \alpha c\lambda^2)\xi_i\right).$$

With $0 \leq \lambda \leq \frac{1}{2\alpha}$, we want $\lambda + \alpha c\lambda^2 \leq \frac{1}{bK_{i-1}}$. This is ensured by $\alpha \geq b\max_{i\in[n]} m_i$.

Define

$$A_k = \left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta\right\}.$$

Then $\omega \in A_k$ implies

$$\begin{aligned}\prod_{i=1}^{k}\frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i} &\leq \exp\left(-(\lambda+\alpha c\lambda^2)\sum_{i=1}^{k}\xi_i + \frac{(\lambda+\alpha c\lambda^2)^2}{2}\sum_{i=1}^{k} aK_{i-1}^2\right)\\ &\leq \exp(-\lambda x + c\lambda^2\beta)\\ &\leq \exp(-\lambda x + 2\lambda^2\beta).\end{aligned}$$

■

Final step. Putting everything together proves the $\theta > 1$ and $\alpha > 0$ case. The rest of the work for the other cases is included in Appendix B. ■

5. Non-convex Convergence

In this section, we prove the following convergence bound.

Theorem 15 *Assume f is L-smooth and that, conditioned on the previous iterates, e_t is centered and $\|e_t\|$ is K-sub-Weibull(θ) with $\theta \geq 1/2$. If $\theta > 1/2$, assume f is ρ-Lipschitz. Let $\delta_1, \delta_2, \delta_3 \in (0,1)$ and define $\delta = \max\{\delta_1, \delta_2, \delta_3\}$. Then, for T iterations of SGD with $\eta_t = c/\sqrt{t+1}$ where $c \leq 1/L$, w.p. $\geq 1 - \delta_1 - \delta_2 - \delta_3$,*

$$\begin{aligned}&\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\frac{1}{\sqrt{t+1}}\|\nabla f(x_t)\|^2\\ &\quad\leq \frac{4(f(x_0)-f^\star)}{c\sqrt{T}} + \frac{\gamma(\theta)\log(1/\delta_3)}{\sqrt{T}} + \frac{4LK^2c(4e\theta\log(2/\delta_1))^{2\theta}\log(T+1)}{\sqrt{T}}\\ &\quad= O\left(\frac{\log(T)\log(1/\delta)^{2\theta} + \gamma(\theta)\log(1/\delta)}{\sqrt{T}}\right)\end{aligned}$$

where

$$\gamma(\theta) = \begin{cases} 64K^2 & \theta = 1/2 \\ 8\max\left\{(4\theta)^\theta eK\rho,\ 4(4\theta)^{2\theta}e^2K^2\right\} & \theta \in (1/2, 1] \\ 8\max\left\{2\log(2T/\delta_2)^{\theta-1}K\rho,\right. & \\ \qquad \left. 4\left[(2^{2\theta+1}+2)\Gamma(2\theta+1) + \dfrac{2^{3\theta}\Gamma(3\theta+1)}{3\log(2T/\delta_2)^{\theta-1}}\right]K^2\right\} & \theta > 1 \end{cases}$$

and observe $\gamma(\theta) = O\left(\log(T/\delta)^{\min\{0,\theta-1\}}\right)$ *for any* $\theta \geq \frac{1}{2}$

Proof As with the PŁ analysis we start the non-convex analysis with Eq. (3). From this, we get a master bound on a weighted sum of the $\|\nabla f(x_t)\|^2$. Our goal is a convergence rate for a weighted average of the $\|\nabla f(x_t)\|^2$ since this would imply convergence to a first-order stationary point, which is the best one can hope for without further assumptions. The master bound is in terms of two sums, an inner product sum and a norm sum. We bound the norm sum using an established sub-Weibull concentration inequality. The inner product sum, on the other hand, is the part that requires the MDS concentration inequality of Theorem 11.

Master bound. As in Section 3, again assume f is L-smooth, hence

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$$

for all $x, y \in \mathbb{R}^d$. Set $y = x_{t+1} = x_t - \eta_t g_t = x_t - \eta_t(\nabla f(x_t) - e_t)$ and $x = x_t$. Then we get

$$f(x_{t+1}) \leq f(x_t) - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla f(x_t)\|^2 + \eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle + \frac{L\eta_t^2}{2}\|e_t\|^2.$$

Summing this and using $f(x_T) \geq f^\star$, we get

$$\sum_{t=0}^{T-1}\eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla f(x_t)\|^2 \leq f(x_0) - f^\star + \underbrace{\sum_{t=0}^{T-1}\eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle}_{\text{inner product sum}} + \underbrace{\frac{L}{2}\sum_{t=0}^{T-1}\eta_t^2\|e_t\|^2}_{\text{norm sum}}. \tag{11}$$

We would like to bound

$$\sum_{t=0}^{T-1}\eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t\rangle \leq O\left(\sum_{t=0}^{T-1}\eta_t^2\|\nabla f(x_t)\|^2\right)$$
$$\text{and } \frac{L}{2}\sum_{t=0}^{T-1}\eta_t^2\|e_t\|^2 \leq O\left(\sum_{t=0}^{T-1}\eta_t^2\right)$$

with high probability so that if $\eta_t = \Theta(1/\sqrt{t+1})$, we get

$$\min_{0\leq t\leq T-1}\|\nabla f(x_t)\|^2 \leq \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\frac{1}{\sqrt{t+1}}\|\nabla f(x_t)\|^2 \leq O\left(\frac{\log(T+1)}{\sqrt{T}}\right)$$

with high probability. While these bounds are in big-O notation, the bounds we prove will be precise.

Norm sum bound. Assume that, conditioned on the previous iterates, e_t is centered and $\|e_t\|$ is K-sub-Weibull(θ) with $\theta \geq 1/2$. Set $\eta_t = c/\sqrt{t+1}$ with $c \leq 1/L$. Let $\delta_1 \in (0,1)$. Using the law of total expectation,

$$\mathbb{E}\left[\exp\left(\left(\frac{\eta_t^2\|e_t\|^2}{\eta_t^2 K^2}\right)^{1/2\theta}\right)\right] \leq 2.$$

Thus, η_t^2 is $\eta_t^2 K$-sub-Weibull(2θ) so we can apply the following sub-Weibull concentration inequality.

Lemma 16 *(Vladimirova et al., 2020, Thm. 1) (Wong et al., 2020, Lma. 5) Suppose $X_1, \ldots, X_n$ are sub-Weibull(θ) with respective parameters $K_1, \ldots, K_n$. Then, for all $\gamma \geq 0$,*

$$P\left(\left|\sum_{i=1}^n X_i\right| \geq \gamma\right) \leq 2\exp\left(-\left(\frac{\gamma}{v(\theta)\sum_{i=1}^n K_i}\right)^{1/\theta}\right),$$

where $v(\theta) = (4e)^\theta$ for $\theta \leq 1$ and $v(\theta) = 2(2e\theta)^\theta$ for $\theta \geq 1$.

Applying Lemma 16, we get, w.p. $\geq 1-\delta_1$,

$$\begin{aligned}\frac{L}{2}\sum_{t=0}^{T-1}\eta_t^2\|e_t\|^2 &\leq LK^2(4e\theta\log(2/\delta_1))^{2\theta}\sum_{t=0}^{T-1}\eta_t^2 \\ &\leq LK^2c^2(4e\theta\log(2/\delta_1))^{2\theta}\log(T+1).\end{aligned}$$

Inner product sum bound. Assume that $\theta > 1$. We will prove the easier cases of $\theta = 1/2$ and $\theta \in (1/2, 1]$ in the appendix. Assume f is ρ-Lipschitz continuous. Define

$$\begin{aligned}\xi_t &= \eta_t(1-L\eta_t)\langle\nabla f(x_t), e_t\rangle \\ K_{t-1} &= \eta_t(1-L\eta_t)K\|\nabla f(x_t)\| \\ \text{and } m_t &= \eta_t(1-L\eta_t)K\rho.\end{aligned}$$

Recall that we defined $\mathcal{F}_t$ as the sigma algebra generated by $e_0, \ldots, e_t$ for all $t \geq 0$ and $\mathcal{F}_{-1} = \{\emptyset, \Omega\}$. So, ξ_t is $\mathcal{F}_t$-measurable and K_{t-1} is $\mathcal{F}_{t-1}$-measurable; hence (ξ_t) and (K_t) are adapted to $(\mathcal{F}_t)$. We also have, for all $t \geq 0$, $0 \leq K_{t-1} \leq m_t$ almost surely, $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$, and

$$\mathbb{E}\left[\exp\left((|\xi_t|/K_{t-1})^{1/\theta}\right) \mid \mathcal{F}_{t-1}\right] \leq 2.$$

In other words, (ξ_t) is a sub-Weibull MDS and (K_t) captures the scale parameters. Let $\delta_2, \delta_3 \in (0,1)$ and define

$$\begin{aligned}\delta &= \delta_2 \\ \beta &= 0 \\ \lambda &= \frac{1}{2\alpha} \\ \text{and } x &= 2\alpha\log(1/\delta_3).\end{aligned}$$

Applying Theorem 11, we get, for all $\alpha \geq bK\rho c$, w.p. $\geq 1 - 2\delta_2 - \delta_3$,

$$\sum_{t=0}^{T-1} \eta_t(1 - L\eta_t)\langle \nabla f(x_t), e_t \rangle \leq 2\alpha \log(1/\delta_3) + \frac{aK^2}{\alpha} \sum_{t=0}^{T-1} \eta_t^2 (1 - L\eta_t)^2 \|\nabla f(x_t)\|^2.$$

Combining this with the norm sum bound and master bound, we get, for all $\alpha \geq bK\rho c$, w.p. $\geq 1 - \delta_1 - 2\delta_2 - \delta_3$,

$$\sum_{t=0}^{T-1} \eta_t \nu_t \|\nabla f(x_t)\|^2 \leq f(x_0) - f^\star + 2\alpha \log(1/\delta_3) + LK^2c^2(4e\theta \log(2/\delta_1))^{2\theta} \log(T+1)$$

where

$$\nu_t = 1 - \frac{L\eta_t}{2} - \frac{aK^2}{\alpha}\eta_t(1 - L\eta_t)^2.$$

We want to bound ν_t away from zero. To do so, assume $c \leq \frac{1}{L}$ and $\alpha \geq 4aK^2c$. Then $\nu_t \geq \frac{1}{4}$. Setting

$$\alpha = \max\{bK\rho, 4aK^2\}c$$

and plugging in a and b completes the proof. ■

Remark 17 *Note that Lipschitz continuity follows immediately if the iterates are bounded. This might lead one to consider using projected SGD, but there are certain issues preventing us from analyzing it, which we discuss in Appendices C and D. But, is Lipschitz continuity even a necessary assumption? Li and Liu (2022), building off of our pre-print, were actually able to relax the Lipschitz continuity assumption to the assumption that* $\frac{1}{\sqrt{t+1}}\|\nabla f(x_t)\| \leq \rho \; \forall t \geq 0$. *To see that this works, note that we can change the definition of* m_t *to* $c(1 - L\eta_t)K\rho$ *and the rest of the analysis, including the result, still holds.*

Remark 18 *Can we extend the analysis beyond sub-Weibull? Yes, but then we would not get logarithmic dependence on* $1/\delta$. *For example, if we assume* $\mathbb{E}\left[\|e_t\|^p \mid \mathcal{F}_{t-1}\right]$ *for some* $p > 2$, *then we could use Corollary 3 instead of Corollary 2 of Bakhshizadeh et al. (2023). This would give us a* $O(\log(1/\delta)/\sqrt{T} + 1/(\delta^a T^b))$ *convergence rate for some* $b > 1/2$. *Thus, we would not have a logarithmic dependence on* $1/\delta$ *in general, but would approach such a dependence as the number of iterations increases.*

6. Post-processing

Note that the results of Theorem 15 are in terms of $\frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \|\nabla f(x_t)\|^2$ which is not a particularly useful quantity by itself. To get a bound in terms of a single iterate, we prove the following probability result, introduce a novel post-processing strategy which outputs a single iterate x, and apply the probability result to bound $\|\nabla f(x)\|^2$ with high probability.

Theorem 19 *Let $T \in \mathbb{N}$. For all $t \in [T]$, let $Z = t$ with probability p_t, where $\sum_{t=1}^{T} p_t = 1$. Let $Z_1, \ldots, Z_n$ be independent copies of Z. Let $Y = \{Z_1, \ldots, Z_n\}$. Let X be an $\mathbb{R}_+^T$-valued random variable independent of Z. Then*

$$P\left(\min_{t \in Y} X_t > e\gamma\right) \leq \exp(-n) + P\left(\sum_{t=1}^{T} p_t X_t > \gamma\right) \ \forall \gamma > 0.$$

Proof First, letting $\gamma > 0$ and $x \in \mathbb{R}_+^T$,

$$\begin{aligned}
P\left(\min_{t \in Y} x_t > \gamma\right) &= P\left(\bigcap_{i=1}^{n} \{x_{Z_i} > \gamma\}\right) \\
&\overset{(i)}{=} \prod_{i=1}^{n} P\left(x_{Z_i} > \gamma\right) \\
&= P\left(x_Z > \gamma\right)^n \\
&\overset{(ii)}{\leq} \left(\frac{1}{\gamma}\mathbb{E}\left[x_Z\right]\right)^n \\
&= \left(\frac{1}{\gamma}\sum_{t=1}^{T} p_t x_t\right)^n,
\end{aligned}$$

where (i) follows by the independence of the Z_i and (ii) follows by Markov's inequality since x_Z is non-negative almost surely. Next, define

$$\begin{aligned}
A &= \left\{ x \in \mathbb{R}_+^T \;\middle|\; \sum_{t=1}^{T} p_t x_t \leq \gamma \right\} \\
B &= \left\{ (x, y) \in \mathbb{R}_+^T \times [T]^n \;\middle|\; x_{y_i} > e\gamma \ \forall i \in [n] \right\}.
\end{aligned}$$

Observe,

$$\begin{aligned}
P\left((X, Y) \in B\right) &\overset{(i)}{=} P\left(X \in A, (X, Y) \in B\right) + P\left(X \in A^c, (X, Y) \in B\right) \\
&\overset{(ii)}{=} \int_A P\left((x, Y) \in B\right) \mu(dx) + \int_{A^c} P\left((x, Y) \in B\right) \mu(dx) \\
&\leq \int_A \left(\frac{1}{e\gamma}\sum_{t=1}^{T} p_t x_t\right)^n \mu(dx) + \int_{A^c} \mu(dx) \\
&\leq \exp(-n) \int_A \mu(dx) + P\left(X \in A^c\right) \\
&\leq \exp(-n) + P\left(\sum_{t=1}^{T} p_t X_t > \gamma\right)
\end{aligned}$$

where (i) follows from the law of total probability and (ii) follows from Theorem 20.3 of Billingsley (1995) since X and Y are independent. ■

Corollary 20 *Assume f is differentiable. Let $\delta_{iter} \in (0,1)$ and $T \in \mathbb{N}$. Set $n_{iter} = \lceil \log(1/\delta_{iter}) \rceil$. Sample n_{iter} indices with replacement from $\{0, \dots, T-1\}$ with probabilities $p_0, \dots, p_{T-1}$ to form the set $S = \{s_1, \dots, s_{n_{iter}}\}$. Then, for T iterations of SGD with any step-size sequence (η_t),*

$$P\left(\min_{t \in S} \|\nabla f(x_t)\|^2 > e\gamma\right) \leq P\left(\sum_{t=0}^{T-1} p_t \|\nabla f(x_t)\|^2 > \gamma\right) + \delta_{iter} \ \forall \gamma > 0.$$

To combine the corollary with Theorem 15, we can set

$$p_t = \frac{1/\sqrt{t+1}}{\sum_{t=0}^{T-1} 1/\sqrt{t+1}} \text{ and use that } \frac{1}{\sum_{t=0}^{T-1} 1/\sqrt{t+1}} \leq \frac{1}{\sqrt{T}}.$$

To see the merit of this post-processing strategy, let's compare it to a naive approach one might take to apply the result of Theorem 15. The standard trick is to observe

$$\min_{0 \leq t \leq T-1} \|\nabla f(x_t)\|^2 \leq \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \|\nabla f(x_t)\|^2. \tag{12}$$

So, we could keep track of $\|\nabla f(x_t)\|$ at every iteration and record the iterate where it is lowest. However, this requires exact gradient information, which may be more costly than the stochastic gradient used in the algorithm. In Ghadimi and Lan (2013), they pick index s with probability proportional to $1/\sqrt{s+1}$ so that $\mathbb{E}\left[\|\nabla f(x_s)\|^2\right]$ is proportional to the right-hand side of Eq. (12). They do this for $\Theta(\log(1/\delta))$ *runs* and pick the best of the runs. Corollary 20, on the other hand, allows us to sample a set S of $n_{\text{iter}} = \Theta(\log(1/\delta))$ indices and pick the best *iterate* from among these samples. Hence, we call δ_{iter} the iterate sampling failure probability.

But, Corollary 20 is not the end of the story since to compute even $\text{argmin}_{t \in S} \|\nabla f(x_t)\|^2$ still requires full gradient information. In the sample average approximation setting, this can be obtained by running on the full batch of data (rather than a mini-batch). However, if this is computationally infeasible or if we are in the stochastic approximation setting, then we instead have to use empirical gradients over a test or validation set. This is what we do for the full post-processing strategy presented in the following theorem.

Theorem 21 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $F : \mathbb{R}^d \times \Omega \to \mathbb{R}$ and assume $F(\cdot; \xi)$ is differentiable for all $\xi \in \Omega$. Let $f = \mathbb{E}[F(\cdot; \xi)]$. Assume $\nabla f = \mathbb{E}[\nabla F(\cdot; \xi)]$ and $\mathbb{E}\left[\|\nabla f(x) - \nabla F(x; \xi)\|^2\right] \leq \sigma^2 \ \forall x \in \mathbb{R}^d$. Let $\delta_{iter}, \delta_{emp} \in (0,1)$ and $T \in \mathbb{N}$. Set $n_{iter} = \lceil \log(1/\delta_{iter}) \rceil$ and $n_{emp} = \lceil 6(n_{iter}+1)\sigma^2/(e\gamma\delta_{emp}) \rceil$. Apply the following procedure:*

1. *Sample n_{iter} indices with replacement from $\{0, \dots, T-1\}$ with probabilities $p_0, \dots, p_{T-1}$ to form the set $S = \{s_1, \dots, s_{n_{iter}}\}$.*
2. *Sample $\xi_1, \dots, \xi_{n_{emp}}$ independently.*
3. *Run T iterations of SGD with any step-size sequence (η_t) to form (x_t).*

4. Compute

$$x = \underset{t \in S}{argmin} \left\| \frac{1}{n_{emp}} \sum_{j=0}^{n_{emp}} \nabla F(x_t; \xi_j) \right\|^2.$$

Then,

$$P\left(\|\nabla f(x)\|^2 > 5e\gamma\right) \leq P\left(\sum_{t=0}^{T-1} p_t \|\nabla f(x_t)\|^2 > \gamma\right) + \delta_{iter} + \delta_{emp} \ \forall \gamma > 0.$$

Proof Apply Theorem 2.4 of Ghadimi and Lan (2013) to get

$$P\left(\|\nabla f(x)\|^2 > 5e\gamma\right) \leq P\left(\min_{t \in S} \|\nabla f(x_t)\|^2 > e\gamma\right) + \frac{6(n_{\text{iter}} + 1)\sigma^2}{e\gamma n_{\text{emp}}} \ \forall \gamma, \lambda > 0.$$

Using the definitions of n_{iter} and n_{emp}, and applying Corollary 20, proves the result. ■

We call δ_{emp} the empirical gradient failure probability. Note that if $\|\nabla f(x) - G(x, \xi)\|$ is K-sub-Weibull(θ), then e_t is centered and $\|e_t\|$ is K-sub-Weibull(θ), in which case both Theorem 15 and Theorem 21 apply, with $\sigma^2 = 2\Gamma(2\theta + 1)K^2$ (by Lemma 6) for the latter. Also, note that for both Corollary 20 and Theorem 21, while we have to specify T in advance, we only have to take $\max S$ iterations to apply the bound from Theorem 15 to the post-processing output.

7. Neural Network Example

Consider the two layer neural network model

$$x \mapsto \phi(x^T W)a$$

where ϕ is a differentiable activation function applied coordinate-wise, $x \in \mathbb{R}^d$ is a data point (feature vector), $W \in \mathbb{R}^{d \times m}$ is the first-layer weights, and $a \in \{\pm 1\}^m$ is the second-layer weights. If we are given a data set $X \in \mathbb{R}^{d \times n}$ and labels $y \in \mathbb{R}^n$, then we can train W (leaving a fixed for simplicity) using the squared loss:

$$f(W) = \frac{1}{2}\|\phi(X^T W)a - y\|^2.$$

In this case,

$$\text{vec}(\nabla f(W)) = \left(\text{diag}(a)\phi'(W^T X) * X\right)\left(\phi(X^T W)a - y\right)$$

where $*$ is the Khatri-Rao product (Oymak and Soltanolkotabi, 2020). For our example, we use the GELU activation, $\phi(x) = x[1 + \text{erf}(x/\sqrt{2})]/2$ (Hendrycks and Gimpel, 2016), which satisfies $|\phi'(x)| \leq 1.13$ and $|\phi''(x)| \leq .11$ for all $x \in \mathbb{R}$. Thus, we can apply Lemma 2. Since $\lim_{x \to \infty} \phi''(x) = 0 = \lim_{x \to -\infty} \phi''(x)$, we heuristically set $b = 0$ in the lemma and estimate the strong smoothness of f as $\approx m\|X\|_2^2$, setting our step-size accordingly.

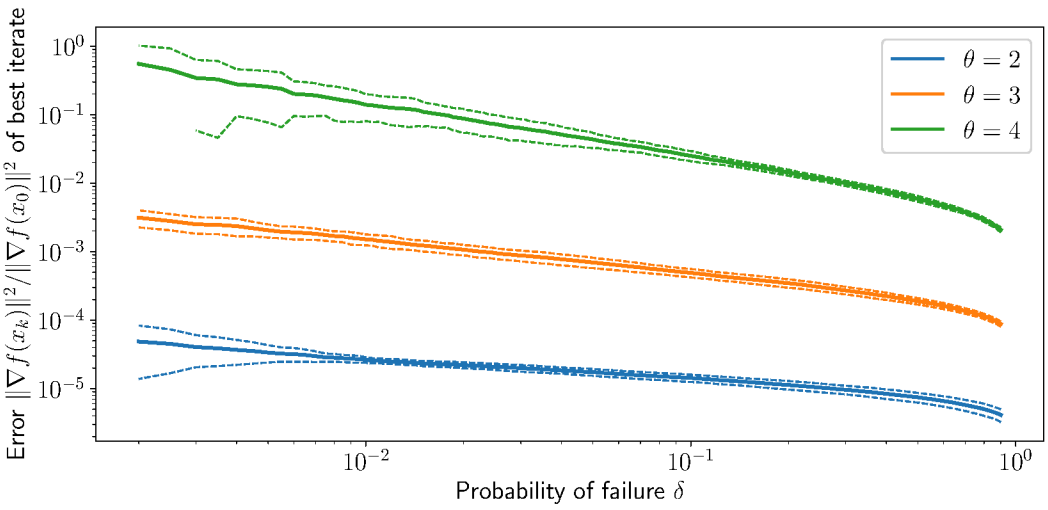

Figure 1: Empirical $1-\delta$ convergence error, averaged over 10000 runs. The dashed lines show the mean $\pm$ one standard deviation (computed over 5 blocks of 2000 runs each). The data are less reliable for small δ.

In order to demonstrate the effect of gradient noise on convergence error, we train W on a fixed synthetic data set while injecting noise into the gradient. The labels come from a neural network model with width, m', larger than the width of the training model. We also make sure the total number of trainable parameters is less than n so that $f^\star > 0$. The noise we inject has uniformly random direction and Weibull norm with scale parameter $K = 1$ and shape parameter $1/\theta$, with $\theta \in \{2, 3, 4\}$. We keep the same initialization for all trials. We run 100 iterations of SGD and then define the convergence error to be the best gradient norm squared divided by the initial gradient norm squared. We compute the empirical CDF of the convergence error over 10000 trials and then consider δ in the ranges $[0.1, 0.2]$, $[0.01, 0.1]$, and $[0.001, 0.01]$. We care about the dependence of the convergence error on δ for small δ, but for too small of δ, the empirical CDF is not a good approximation to the true CDF (see Fig. 1) due to the nature of order statistics. Our code can be found at `https://github.com/liammadden/sgd`.

From Theorem 15, for a particular range of δ (and for fixed T), the upper bound has dependence either $\log(1/\delta)^\theta$, $\log(1/\delta)$, or $\log(1/\delta)^{2\theta}$. For sufficiently small δ, the $\log(1/\delta)^{2\theta}$ will dominate, but the upper limit of this range of δ may be smaller than the lower limit of the range of δ for which the empirical CDF is a good approximation to the true CDF. In other words, if we showed that the true CDF for this particular example has $\log(1/\delta)^{2\theta}$ dependence for δ sufficiently small, then we would have shown that the δ-dependence of the upper bound in Theorem 15 is tight. However, with the empirical CDF, we are only able to show that the exponent increases as δ shrinks. Figure 1 shows the dependence for the three different ranges. We assume the convergence error has dependence $b\log(1/\delta)^a$ and find the line of best fit as $\log(b) + a\log\log(1/\delta)$. The different values of a are given in Table 1.

a	$\delta \in [0.1, 0.2]$	$\delta \in [0.01, 0.1]$	$\delta \in [0.001, 0.01]$
$\theta = 2$	0.64	0.85	1.40
$\theta = 3$	0.93	1.56	2.33
$\theta = 4$	1.51	2.38	4.08

Table 1: Empirically estimated exponents of $\log(1/\delta)$

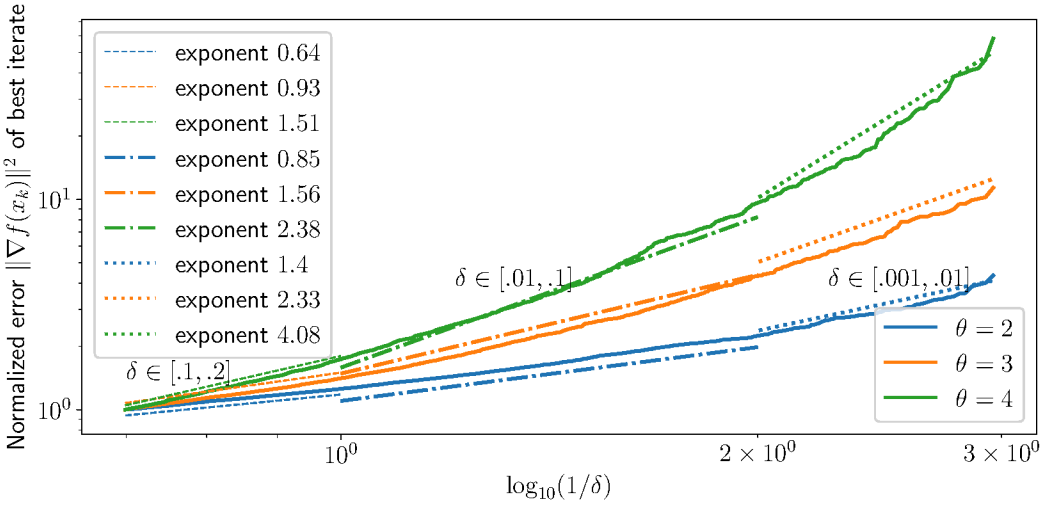

Figure 2: Same data as Fig. 1 but each line series is normalized, and the x-axis is $\log(1/\delta)$ *and* plotted on a logarithmic scale, so $\log(1/\delta)^a$ dependence shows us a straight line with slope a. The δ range is from 0.2 (left side) to 0.01 (right side), since any smaller δ has unreliable statistics. Lines of best fit using the exponents from Table 1 are shown (with arbitrary shifts for clarity).

Our experiments suggest that injected Weibull noise results in convergence error with dependence $\Omega(\log(1/\delta)^{c(\delta)\theta+d(\delta)})$ where $c(\delta)$ increases as δ decreases, thus roughly corroborating our upper bound. In particular, by computing a line of best fit for the a values in each of the three δ ranges, we can estimate that $c(\delta)$ increases from 0.43, to 0.77, to 1.34 and $d(\delta)$ decreases from -0.28, to -0.7, to -1.42, suggesting that we are in the $\log(1/\delta)^{\theta-1}$ regime.

8. Conclusions

This paper analyzed the convergence of SGD for objective functions that satisfy the PŁ condition and for generic non-convex objectives. Under a sub-Gaussian assumption on the gradient error, we showed a high probability convergence rate matching the mean convergence rate for PŁ objectives. Under a sub-Weibull assumption on the noise, we showed a high probability convergence rate matching the mean convergence rate for non-convex

objectives. We also provided and analyzed a post-processing method for choosing a single iterate. To prove the convergence rate, we first proved a Freedman-type inequality for martingale difference sequences that extends previous Freedman-type inequalities beyond the sub-exponential threshold to allow for sub-Weibull tail-decay. Finally, we considered a synthetic neural net problem and showed that the heaviness of the tail of the gradient noise has a direct effect on the heaviness of the tail of the convergence error.

Acknowledgments

All three authors gratefully acknowledge support from the National Science Foundation (NSF), Division of Mathematical Sciences, through the award # 1923298.

Appendix A. Proof of Lemma 1

Here we prove Lemma 1.

Proof First,

$$f(W) = \sum_{i=1}^{n} \frac{1}{2} \left[\phi(x_i^\top W)v - y_i \right]^2$$

and

$$\frac{\partial}{\partial w_{s,t}} \left[\phi(x_i^\top W)v - y_i \right] = \frac{\partial}{\partial w_{s,t}} \sum_{j=1}^{m} v_j \phi(x_i^\top w_j) = v_t \phi'(x_i^T w_t) x_{s,i}.$$

So,

$$\frac{\partial}{\partial w_{s,t}} f(W) = \sum_{i=1}^{n} \left[\phi(x_i^\top W)v - y_i \right] v_t \phi'(x_i^T w_t) x_{s,i}.$$

Thus,

$$\begin{aligned} \frac{\partial^2}{\partial w_{\ell,r} \partial w_{s,t}} f(W) = &\sum_{i=1}^{n} v_r \phi'(x_i^\top w_r) x_{\ell,i} v_t \phi'(x_i^\top w_t) x_{s,i} \\ &+ \delta_{r,t} \sum_{i=1}^{n} \left[\phi(x_i^\top W)v - y_i \right] v_t \phi''(x_i^\top w_t) x_{\ell,i} x_{s,i} \end{aligned}$$

where δ is the Kronecker delta. Let b be the largest absolute element of X. Then, viewing W as a vector,

$$\|\nabla f(W)\|_2 \leq \sqrt{md} \|\nabla f(W)\|_\infty \leq \sqrt{md} nab \|v\|_\infty (a\sqrt{m}\|v\|_2 + \|y\|_\infty)$$

and

$$\|\nabla^2 f(W)\|_2 \leq md \|\nabla^2 f(W)\|_{\max} \leq mdn \left(a^2 b^2 \|v\|_\infty^2 + ab^2 \|v\|_\infty (a\sqrt{m}\|v\|_2 + \|y\|_\infty) \right),$$

proving the result. ■

And here we prove Lemma 2.

Proof Define $F : w \mapsto \phi(X^\top \mathrm{vec}^{-1}(w))v$ and $\mathcal{L} : w \mapsto \|F(w) - y\|_2^2/2$. Then $\|DF(w)\|_2 \leq \sqrt{m}a\|v\|_\infty \|X\|_2$ and $\|DF(w) - DF(u)\|_2 \leq b\|v\|_\infty \|X\|_2 \|X\|_{1,2}\|w - u\|_2$ for all $w, u \in \mathbb{R}^{md}$ by Lemmas 3 and 5 of Oymak and Soltanolkotabi (2020). Let $\rho = \sqrt{m}a\|v\|_\infty \|X\|_2$ and $L = b\|v\|_\infty \|X\|_2 \|X\|_{1,2}$. Observe

$$\|\nabla \mathcal{L}(w)\|_2 = \|DF(w)^\top (F(w) - y)\|_2 \leq \|DF(w)\|_2 \|F(w) - y\|_2 \leq \rho\sqrt{2\alpha}$$

and

$$\begin{aligned}\|\nabla \mathcal{L}(w) - \nabla \mathcal{L}(u)\| &= \|DF(w)^\top (F(w) - F(v))\|_2 + \|(DF(w) - DF(v))^\top (F(v) - y)\|_2 \\ &\leq \|DF(w)\|_2 \|F(w) - F(v)\|_2 + \|DF(w) - DF(v)\|_2 \|F(v) - y\|_2 \\ &\leq \rho^2 \|w - v\|_2 + L\|w - v\|_2 \sqrt{2\alpha} = (\rho^2 + L\sqrt{2\alpha})\|w - v\|_2,\end{aligned}$$

proving the result. ■

Appendix B. Remaining Work for Proof of Theorem 11

First we need the following two lemmas. Lemma 23 extends Proposition 2.7.1 of Vershynin (2018) to interpolate between the sub-Gaussian and sub-exponential regimes.

Lemma 22 *(Vershynin, 2018, Prop. 2.5.2(e)) If X is centered and K-sub-Guassian then* $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp(\lambda^2 K^2)$ $\forall \lambda \in \mathbb{R}$.

Lemma 23 *If X is centered and K-sub-Weibull(θ) with $\theta \in (1/2, 1]$, then* $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2}{2}(4\theta)^{2\theta} e^2 K^2\right)$ *for all* $\lambda \in \left[0, \frac{1}{(4\theta)^\theta eK}\right]$.

Proof First, using Lemma 6 and $\Gamma(x+1) \leq x^x \ \forall x \geq 1$, we can get $\|X\|_p \leq (2\theta)^\theta K p^\theta$ for all $p \geq 1/\theta$, and so, in particular, for all $p \geq 2$. Thus, if $\lambda \in \left[0, \frac{1}{(4\theta)^\theta eK}\right]$, then

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda X)\right] &= \mathbb{E}\left[1 + \lambda X + \sum_{p=2}^{\infty} \frac{(\lambda X)^p}{p!}\right] \\
&= 1 + \sum_{p=2}^{\infty} \frac{\lambda^p \|X\|_p^p}{p!} \\
&\leq 1 + \sum_{p=2}^{\infty} \frac{\lambda^p (2\theta)^{\theta p} K^p p^{\theta p}}{p!} \\
&\leq 1 + \sum_{p=2}^{\infty} \left(\frac{\lambda (2\theta)^\theta eK}{p^{1-\theta}}\right)^p \\
&\leq 1 + \sum_{p=2}^{\infty} \left(\lambda (4\theta)^\theta (e/2) K\right)^p \\
&\leq 1 + \frac{\left(\lambda (4\theta)^\theta (e/2) K\right)^2}{1 - \lambda (4\theta)^\theta (e/2) K} \\
&\leq 1 + 2\left(\lambda (4\theta)^\theta (e/2) K\right)^2 \\
&\leq \exp\left(\frac{\lambda^2}{2} (4\theta)^{2\theta} e^2 K^2\right),
\end{aligned}$$

completing the proof. ■

Then, the next three lemmas allow us to include previous results as special cases of the theorem.

Lemma 24 *(Fan et al., 2015, Thm. 2.6) Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$. For all $i \in [n]$, assume $0 \leq K_{i-1} \leq m_i$ almost surely, $\mathbb{E}\left[\xi_i \mid \mathcal{F}_{i-1}\right] = 0$, and*

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} a K_{i-1}^2\right) \ \forall \lambda \in \left[0, \frac{1}{b K_{i-1}}\right].$$

Then, for all $x, \beta \geq 0$, and $\lambda \in \left[0, \frac{1}{b \max_{i \in [r]} m_i}\right]$,

$$P\left(\bigcup_{k \in [n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} a K_{i-1}^2 \leq \beta \right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right).$$

Proof Define

$$\psi_i = \exp\left(\lambda \xi_i\right)$$

and

$$A_k = \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta \right\}.$$

Then $\omega \in A_k$ implies

$$\begin{aligned}\prod_{i=1}^{k} \frac{\mathbb{E}\left[\psi_i \mid \mathcal{F}_{i-1}\right]}{\psi_i} &\leq \exp\left(-\lambda \sum_{i=1}^{k} \xi_i + \frac{\lambda^2}{2} \sum_{i=1}^{k} aK_{i-1}^2\right) \\ &\leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right).\end{aligned}$$

■

Lemma 25 *(Harvey et al., 2019, Thm. 3.3) Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$. For all $i \in [n]$, assume $K_{i-1} \geq 0$ almost surely, $\mathbb{E}\left[\xi_i \mid \mathcal{F}_{i-1}\right] = 0$, and*

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} a K_{i-1}^2\right) \ \forall \lambda \geq 0.$$

Then, for all $x, \beta, \alpha \geq 0$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k \in [n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha \sum_{i=1}^{k} \xi_i + \beta \right\}\right) \leq \exp(-\lambda x + 2\lambda^2 \beta).$$

Lemma 26 *(Freedman, 1975) Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. For all $i \in [n]$, assume $K_{i-1} \geq 0$ almost surely, $\mathbb{E}\left[\xi_i \mid \mathcal{F}_{i-1}\right] = 0$, and*

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} a K_{i-1}^2\right) \ \forall \lambda \geq 0.$$

Then, for all $x, \beta \geq 0$, and $\lambda \geq 0$,

$$P\left(\bigcup_{k \in [n]} \left\{ \sum_{i=1}^{k} \xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta \right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta\right).$$

If the ξ_i's are sub-Gaussian, that is, if $\theta = 1/2$, then, from Lemma 22,

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} 2K_{i-1}^2\right) \ \forall \lambda \in \mathbb{R},$$

so we can apply Lemma 25 if $\alpha > 0$ or Lemma 26 if $\alpha = 0$.

If the ξ_i's are at most sub-exponential, that is, if $1/2 < \theta \leq 1$, then, from Lemma 23,

$$\mathbb{E}\left[\exp(\lambda \xi_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} (4\theta)^{2\theta} e^2 K_{i-1}^2\right) \ \forall \lambda \in \left[0, \frac{1}{(4\theta)^\theta e K_{i-1}}\right],$$

so we can apply Lemma 14 if $\alpha > 0$ or Lemma 24 if $\alpha = 0$.

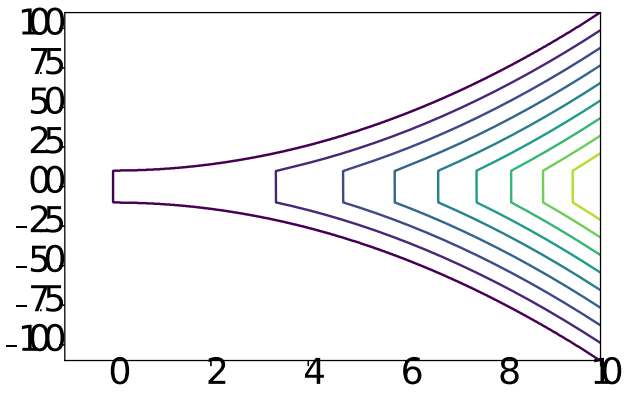

Figure 3: Contour plot of the PŁ function counter-example to projected gradient flow

Appendix C. PŁ Projected SGD

When optimizing over a constraint set, $X \subsetneq \mathbb{R}^d$, if f is strongly convex, then so is f plus the indicator function of X, and results for gradient descent methods easily extend to projected gradient descent. On the other hand, if f is PŁ, then f plus the indicator function is not KL (Kurdyka-Łojasiewicz). This has real impacts on gradient descent algorithms, where gradient descent might converge while projected gradient descent does not. For example, there is a smooth function, a mollified version of $f(x,y) = \left(a(x)_+^2 - b\left(|y| - c\right)_+\right)_+$, such that the PŁ inequality is satisfied but projected gradient descent does *not* converge to a minimizer; we formalize this in the remark below.

Remark 27 *Consider* $f(x,y) = \left(a(x)_+^2 - b\left(|y| - c\right)_+\right)_+$ *where* $(\cdot)_+$ *denotes* $\max(\cdot, 0)$ *and* $a, b > 0, c \geq 0$. *The minimum of* f *is 0 and* $X^\star = \{(x,y) \mid x \leq 0 \text{ or } |y| \geq \frac{a}{b}x^2 + c\}$. *If we use* $\varphi(x) = \mathcal{X}_{B_1(0)} \cdot \exp\left(-1/(1 - \|x\|^2)\right)/\Phi$—*where* $\mathcal{X}$ *denotes the indicator function,* $B_1(0)$ *denotes the ball of radius 1 centered at 0, and* Φ *is the normalization constant—to mollify* f, *then, for* $\epsilon < c$, f_ϵ *has PŁ constant* $2a$ *and smoothness constant* $2a$. *Consider the starting point* $(d, 0)$. *For* a, b, c, d *chosen appropriately, the distance from* $(d, 0)$ *to its projection onto* $X^\star$ *is strictly less than* d. *Thus, if we let* X *be the ball centered at* $(d, 0)$ *with radius equal to exactly that distance, then the constrained problem and the unconstrained problem have the same minimum. However, projected gradient flow, starting from* $(d, 0)$, *ends up stuck at a non-minimizer: the point of* X *closest to* $(0, 0)$. *See Figure 3 for the contour plot when* $a = 1/10$, $b = 1 = c$, *and* $d = 10$.

In order to generalize gradient methods to projected gradient methods under PL-like assumptions, the proper generalization is that the function should satisfy a *proximal* PŁ inequality (Karimi et al., 2016). However, such an assumption is quite restrictive compared to the PŁ inequality. We leave the problem of convergence with just the PŁ inequality, via added noise or a Frank-Wolfe type construction, as a future research direction.

Appendix D. Non-convex Projected SGD

Consider $x_{t+1} = \text{proj}(x_t - \eta_t g_t) = \text{proj}(x_t - \eta_t(\nabla f(x_t) - e_t))$. Define

$$G_t = \frac{x_t - \text{proj}(x_t - \eta_t \nabla f(x_t))}{\eta_t}$$
$$E_t = \frac{x_{t+1} - \text{proj}(x_t - \eta_t \nabla f(x_t))}{\eta_t}.$$

Note that if proj $= I$, then $G_t = \nabla f(x_t)$ and $E_t = e_t$. Moreover, $x_t = \text{proj}(x_t)$ and $x_{t+1} = \text{proj}(x_{t+1})$ so, by the non-expansiveness of proj, $\|G_t\| \leq \|\nabla f(x_t)\|$ and $\|E_t\| \leq \|e_t\|$. We can get even tighter bounds using the second prox theorem (Beck, 2017, Thm. 6.39): $\|G_t\|^2 \leq \langle \nabla f(x_t), G_t \rangle$ and $\langle G_t, E_t \rangle \leq \langle \nabla f(x_t), E_t \rangle$.

It is easy to come up with an example where $\|\nabla f(x_t)\|$ does not go to zero, so we would like to bound $\|G_t\|$ instead. We start as usual:

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

Focusing on the norm term,

$$\frac{L}{2}\|x_{t+1} - x_t\|^2 = \frac{L\eta_t^2}{2}\|G_t\|^2 - L\eta_t^2 \langle G_t, E_t \rangle + \frac{L\eta_t^2}{2}\|E_t\|^2.$$

Focusing on the inner product term,

$$\begin{aligned}
\langle \nabla f(x_t), x_{t+1} - x_t \rangle &= \eta_t \langle \nabla f(x_t), E_t - G_t \rangle \\
&= \eta_t \langle \nabla f(x_t), E_t \rangle - \eta_t \langle \nabla f(x_t), G_t \rangle \\
&\not\leq \eta_t \langle G_t, E_t \rangle - \eta_t \|G_t\|^2.
\end{aligned}$$

Unfortunately, we cannot proceed any further. Ghadimi et al. (2016) are able to get around this but at the cost of getting $\sum_{t=0}^{T-1} \eta_t \|e_t\|^2 = O(\sqrt{T})$ instead of $\sum_{t=0}^{T-1} \eta_t^2 \|e_t\|^2 = O(\log(T))$. To mitigate this, they require an *increasing* batch-size. Reddi et al. (2016) were able to remove this requirement, but only for non-convex projected SVRG *not* non-convex projected SGD. Thus, we leave the analysis of non-convex projected SGD as an open problem.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Neural Information Processing Systems (NeurIPS)*, volume 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, volume 97, pages 242–252, 2019b.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

Milad Bakhshizadeh, Arian Maleki, and Victor H. de la Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3): 1655–1685, 2023.

Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization, 2017.

Dimitri Bertsekas and John Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 3 edition, 1995.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Neural Information Processing Systems (NeurIPS)*, volume 20, 2008.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning (ICML)*, volume 80, pages 745–754, 2018.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97, pages 5827–5837, 2019.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Neural Information Processing Systems (NeurIPS)*, volume 34, pages 4883–4895, 2021.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97, pages 1675–1685, 2019.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

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

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, volume 139, pages 3964–3975, 2021.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory (COLT)*, volume 99, pages 1579–1613, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning (ICML)*, volume 162, pages 12931–12963, 2022.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. In *Workshop on "Beyond first-order methods in ML systems" at the 37th Internation Conference on Machine Learning (ICML)*, 2020.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Liam Madden and Christos Thrampoulidis. Memory capacity of two layer neural networks with smooth activations. *SIAM Journal on Mathematics of Data Science*, 6(3):679–702, 2024.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Arkadi Nemirovsky and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Switzerland, 2 edition, 2018.

Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. In *Neural Information Processing Systems (NeurIPS)*, pages 12589–12601, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. In *Science meets Engineering of Deep Learning (SEDL) workshop, 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Vivak Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions. *Mathematical Programming*, 164, 2021.

Vivak Patel, Shushu Zhang, and Bowen Tian. Global convergence and stability of stochastic gradient descent. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 36014–36025, 2022.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Neural Information Processing Systems (NeurIPS)*, pages 1153–1161, 2016.

Kevin Scaman and Cedric Malherbe. Robustness analysis of non-convex stochastic gradient descent using biased expectations. In *Neural Information Processing Systems (NeurIPS)*, volume 33, pages 16377–16387, 2020.

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory (COLT)*, pages 3935–3971, 2021.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

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 (ICLR)*, 2020.