The Heavy-Tail Phenomenon in SGD

Mert Gürbüzbalaban [1] **Umut Şimşekli** [2] **Lingjiong Zhu** [3]

Abstract

In recent years, various notions of capacity and complexity have been proposed for characterizing the generalization properties of stochastic gradient descent (SGD) in deep learning. Some of the popular notions that correlate well with the performance on unseen data are (i) the 'flatness' of the local minimum found by SGD, which is related to the eigenvalues of the Hessian, (ii) the ratio of the stepsize η to the batch-size b, which essentially controls the magnitude of the stochastic gradient noise, and (iii) the 'tail-index', which measures the heaviness of the tails of the network weights at convergence. In this paper, we argue that these three seemingly unrelated perspectives for generalization are deeply linked to each other. We claim that depending on the structure of the Hessian of the loss at the minimum, and the choices of the algorithm parameters η and b, the distribution of the SGD iterates will converge to a *heavy-tailed* stationary distribution. We rigorously prove this claim in the setting of quadratic optimization: we show that even in a simple linear regression problem with independent and identically distributed data whose distribution has finite moments of all order, the iterates can be heavy-tailed with infinite variance. We further characterize the behavior of the tails with respect to algorithm parameters, the dimension, and the curvature. We then translate our results into insights about the behavior of SGD in deep learning. We support our theory with experiments conducted on synthetic data, fully connected, and convolutional neural networks.

[1]Department of Management Science and Information Systems, Rutgers Business School, Piscataway, USA [2] INRIA - Département d'Informatique de l'École Normale Supérieure - PSL Research University, Paris, France [3]Department of Mathematics, Florida State University, Tallahassee, USA. Correspondence to: Mert Gürbüzbalaban <mg1366@rutgers.edu>, Umut Şimşekli <umut.simsekli@inria.fr>, Lingjiong Zhu <zhu@math.fsu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

The learning problem in neural networks can be expressed as an instance of the well-known *population risk minimization* problem in statistics, given as follows:

$$\min_{x\in\mathbb{R}^d} F(x) := \mathbb{E}_{z\sim\mathcal{D}}[f(x,z)], \tag{1.1}$$

where $z \in \mathbb{R}^p$ denotes a random data point, $\mathcal{D}$ is a probability distribution on $\mathbb{R}^p$ that denotes the law of the data points, $x \in \mathbb{R}^d$ denotes the parameters of the neural network to be optimized, and $f : \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}_+$ denotes a measurable cost function, which is often non-convex in x. While this problem cannot be attacked directly since $\mathcal{D}$ is typically unknown, if we have access to a *training dataset* $S = \{z_1, \dots, z_n\}$ with n independent and identically distributed (i.i.d.) observations, i.e., $z_i \sim_{\text{i.i.d.}} \mathcal{D}$ for $i = 1, \dots, n$, we can use the *empirical risk minimization* strategy, which aims at solving the following optimization problem (Shalev-Shwartz & Ben-David, 2014):

$$\min_{x\in\mathbb{R}^d} f(x) := f(x,S) := (1/n)\sum\nolimits_{i=1}^{n} f^{(i)}(x), \tag{1.2}$$

where $f^{(i)}$ denotes the cost induced by the data point z_i. The stochastic gradient descent (SGD) algorithm has been one of the most popular algorithms for addressing this problem:

$$x_k = x_{k-1} - \eta\nabla\tilde{f}_k(x_{k-1}), \tag{1.3}$$
$$\text{where} \quad \nabla\tilde{f}_k(x) := (1/b)\sum\nolimits_{i\in\Omega_k} \nabla f^{(i)}(x).$$

Here, k denotes the iterations, $\eta > 0$ is the stepsize (also called the learning-rate), $\nabla\tilde{f}$ is the stochastic gradient, b is the batch-size, and $\Omega_k \subset \{1, \dots, n\}$ is a random subset with $|\Omega_k| = b$ for all k.

Even though the practical success of SGD has been proven in many domains, the theory for its generalization properties is still in an early phase. Among others, one peculiar property of SGD that has not been theoretically well-grounded is that, depending on the choice of η and b, the algorithm can exhibit significantly different behaviors in terms of the performance on unseen test data.

A common perspective over this phenomenon is based on the 'flat minima' argument that dates back to Hochreiter & Schmidhuber (1997), and associates the performance with

the 'sharpness' or 'flatness' of the minimizers found by SGD, where these notions are often characterized by the magnitude of the eigenvalues of the Hessian, larger values corresponding to sharper local minima (Keskar et al., 2017). Recently, Jastrzębski et al. (2017) focused on this phenomenon as well and empirically illustrated that the performance of SGD on unseen test data is mainly determined by the stepsize η and the batch-size b, i.e., larger η/b yields better generalization. Revisiting the flat-minima argument, they concluded that the ratio η/b determines the flatness of the minima found by SGD; hence the difference in generalization. In the same context, Şimşekli et al. (2019b) focused on the statistical properties of the gradient noise $(\nabla \tilde{f}_k(x) - \nabla f(x))$ and illustrated that under an isotropic model, the gradient noise exhibits a heavy-tailed behavior, which was also confirmed in follow-up studies (Zhang et al., 2020; Zhou et al., 2020). Based on this observation and a metastability argument (Pavlyukevich, 2007), they showed that SGD will 'prefer' wider basins under the heavy-tailed noise assumption, without an explicit mention of the cause of the heavy-tailed behavior. More recently, Xie et al. (2021) studied SGD with anisotropic noise and showed with a density diffusion theory approach that it favors flat minima.

In another recent study, Martin & Mahoney (2019) introduced a new approach for investigating the generalization properties of deep neural networks by invoking results from heavy-tailed random matrix theory. They empirically showed that the eigenvalues of the weight matrices in different layers exhibit a *heavy-tailed* behavior, which is an indication that the weight matrices themselves exhibit heavy tails as well (Ben Arous & Guionnet, 2008). Accordingly, they fitted a power law distribution to the empirical spectral density of individual layers and illustrated that heavier-tailed weight matrices indicate better generalization. Very recently, Şimşekli et al. (2020) formalized this argument in a mathematically rigorous framework and showed that such a heavy-tailed behavior diminishes the 'effective dimension' of the problem, which in turn results in improved generalization. While these studies form an important initial step towards establishing the connection between heavy tails and generalization, the *originating cause* of the observed heavy-tailed behavior is yet to be understood.

Contributions. In this paper, we argue that these three seemingly unrelated perspectives for generalization are deeply linked to each other. We claim that, depending on the choice of the algorithm parameters η and b, the dimension d, and the curvature of f (to be precised in Section 3), SGD exhibits a 'heavy-tail phenomenon', meaning that the law of the iterates converges to a heavy-tailed distribution. We rigorously prove that, this phenomenon is not specific to deep learning and in fact it can be observed even in surprisingly simple settings: we show that when f is chosen as a simple quadratic function and the data points are i.i.d. from a continuous distribution supported on $\mathbb{R}^d$ with light tails, the distribution of the iterates can still converge to a heavy-tailed distribution with arbitrarily heavy tails, hence with infinite variance. If in addition, the input data is isotropic Gaussian, we are able to provide a sharp characterization of the tails where we show that (i) the tails become *monotonically heavier* for increasing curvature, increasing η, or decreasing b, hence relating the heavy-tails to the ratio η/b and the curvature, (ii) the law of the iterates converges exponentially fast towards the stationary distribution in the Wasserstein metric, (iii) there exists a higher-order moment (e.g., variance) of the iterates that diverges *at most* polynomially-fast, depending on the heaviness of the tails at stationarity. More generally, if the input data is not Gaussian, our monotonicity results extend where we can show that a lower bound on the thickness of the tails (which will be defined formally in Section 3) is monotonic with respect to η, b, d and the curvature. To the best of our knowledge, these results are the first of their kind to rigorously characterize the empirically observed heavy-tailed behavior of SGD with respect to the parameters η, b, d, and the curvature, with explicit convergence rates.[1] Finally, we support our theory with experiments conducted on both synthetic data and neural networks. Our experimental results provide strong empirical support that our theory extends to deep learning settings for both fully connected and convolutional networks.

2. Technical Background

Heavy-tailed distributions with a power-law decay. A real-valued random variable X is said to be *heavy-tailed* if the right tail or the left tail of the distribution decays slower than any exponential distribution. We say X has heavy (right) tail if $\lim_{x\to\infty} \mathbb{P}(X \geq x)e^{cx} = \infty$ for any $c > 0$. [2]

[1] We note that in a concurrent work, which very recently appeared on arXiv, Hodgkinson & Mahoney (2020) showed that heavy tails with power laws arise in more general Lipschitz stochastic optimization algorithms that are contracting on average for strongly convex objectives near infinity with positive probability. Our Theorem 2 and Lemma 16 are more refined as we focus on the special case of SGD for linear regression, where we are able to provide constants which *explicitly* determine the tail-index as an expectation over data and SGD parameters (see also eqn. (3.9)). Due to the generality of their framework, Theorem 1 in Hodgkinson & Mahoney (2020) is more implicit and it cannot provide such a characterization of these constants, however it can be applied to other algorithms beyond SGD. All our other results (including Theorem 4 – monotonicity of the tail-index and Corollary 11 – central limit theorem for the ergodic averages) are all specific to SGD and cannot be obtained under the framework of Hodgkinson & Mahoney (2020). We encourage the readers to refer to Hodgkinson & Mahoney (2020) for the treatment of more general stochastic recursions.

[2] A real-valued random variable X has heavy (left) tail if $\lim_{x\to\infty} \mathbb{P}(X \leq -x)e^{c|x|} = \infty$ for any $c > 0$.

Similarly, an $\mathbb{R}^d$-valued random vector X has heavy tail if $u^T X$ has heavy right tail for some vector $u \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\| = 1\}$ is the unit sphere in $\mathbb{R}^d$.

Heavy tail distributions include α-stable distributions, Pareto distribution, log-normal distribution and the Weibull distribution. One important class of the heavy-tailed distributions is the distributions with *power-law* decay, which is the focus of our paper. That is, $\mathbb{P}(X \geq x) \sim c_0 x^{-\alpha}$ as $x \to \infty$ for some $c_0 > 0$ and $\alpha > 0$, where $\alpha > 0$ is known as the *tail-index*, which determines the tail thickness of the distribution. Similarly, we say that the random vector X has power-law decay with tail-index α if for some $u \in \mathbb{S}^{d-1}$, we have $\mathbb{P}(u^T X \geq x) \sim c_0 x^{-\alpha}$, for some $c_0, \alpha > 0$.

Stable distributions. The class of α-stable distributions are an important subclass of heavy-tailed distributions with a power-law decay, which appears as the limiting distribution of the generalized CLT for a sum of i.i.d. random variables with infinite variance (Lévy, 1937). A random variable X follows a symmetric α-stable distribution denoted as $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ if its characteristic function takes the form:

$$\mathbb{E}\left[e^{itX}\right] = \exp\left(-\sigma^\alpha |t|^\alpha\right), \qquad t \in \mathbb{R},$$

where $\sigma > 0$ is the scale parameter that measures the spread of X around 0, and $\alpha \in (0, 2]$ is known as the tail-index, and $\mathcal{S}\alpha\mathcal{S}$ becomes heavier-tailed as α gets smaller. The probability density function of a symmetric α-stable distribution, $\alpha \in (0, 2]$, does not yield closed-form expression in general except for a few special cases. When $\alpha = 1$ and $\alpha = 2$, $\mathcal{S}\alpha\mathcal{S}$ reduces to the Cauchy and the Gaussian distributions, respectively. When $0 < \alpha < 2$, α-stable distributions have their moments being finite only up to the order α in the sense that $\mathbb{E}[|X|^p] < \infty$ if and only if $p < \alpha$, which implies infinite variance.

Wasserstein metric. For any $p \geq 1$, define $\mathcal{P}_p(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^d$ with the finite p-th moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1, \nu_2 \in \mathcal{P}_p(\mathbb{R}^d)$, we define the standard p-Wasserstein metric (Villani, 2009):

$$\mathcal{W}_p(\nu_1, \nu_2) := \left(\inf \mathbb{E}\left[\|Z_1 - Z_2\|^p\right]\right)^{1/p},$$

where the infimum is taken over all joint distributions of the random variables Z_1, Z_2 with marginal distributions ν_1, ν_2.

3. Setup and Main Theoretical Results

We first observe that SGD (1.3) is an iterated random recursion of the form $x_k = \Psi(x_{k-1}, \Omega_k)$, where the map $\Psi : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^d$, $\mathcal{S}$ denotes the set of all subsets of $\{1, 2, \ldots, N\}$ and Ω_k is random and i.i.d. If we write $\Psi_\Omega(x) = \Psi(x, \Omega)$ for notational convenience where Ω has the same distribution as Ω_k, then Ψ_Ω is a random map and

$$x_k = \Psi_{\Omega_k}(x_{k-1}). \tag{3.1}$$

Such random recursions are studied in the literature. If this map is Lipschitz on average, i.e.

$$\mathbb{E}[L_\Omega] < \infty, \text{ with } L_\Omega := \sup_{x,y \in \mathbb{R}^d} \frac{\|\Psi_\Omega(x) - \Psi_\Omega(y)\|}{\|x - y\|}, \tag{3.2}$$

and is mean-contractive, i.e. if $\mathbb{E}\log(L_\Omega) < 0$ then it can be shown under further technical assumptions that the distribution of the iterates converges to a unique stationary distribution x_∞ geometrically fast (the Prokhorov distance is proportional to ρ^k for some $\rho < 1$) although the rate of convergence ρ is not explicitly known in general (Diaconis & Freedman, 1999). However, much less is known about the tail behavior of the limiting distribution x_∞ except when the map $\Psi_\Omega(x)$ has a linear growth for large x. The following result characterizes the tail-index under such assumptions for dimension $d = 1$. We refer the readers to Mirek (2011) for general d.

Theorem 1. *(Mirek (2011), see also Buraczewski et al. (2016)) Assume stationary solution to (3.1) exists and:*

(i) There exists a random variable $M(\Omega)$ and a random variable $B(\Omega) > 0$ such that for a.e. Ω, $\|\Psi_\Omega(x) - M(\Omega)x\| \leq B(\Omega)$ for every x;

(ii) The conditional law of $\log \|M(\Omega)\|$ given $M(\Omega) \neq 0$ is non-arithmetic;

(iii) There exists $\alpha > 0$ such that $\mathbb{E}\|M(\Omega)\|^\alpha = 1$, $\mathbb{E}\|B(\Omega)\|^\alpha < \infty$ and $\mathbb{E}[\|M(\Omega)\|^\alpha \log^+ \|M(\Omega)\|] < \infty$ where $\log^+(x) := \max(\log(x), 0)$.

Then, it holds that $\lim_{t\to\infty} t^\alpha \mathbb{P}(\|x_\infty\| > t) = c_0$ for some constant $c_0 > 0$.

Relaxations of the assumptions of Theorem 1 which require only lower and upper bounds on the growth of Ψ_Ω have also been recently developed (Hodgkinson & Mahoney, 2020; Alsmeyer, 2016). Unfortunately, it is highly non-trivial to verify such assumptions in practice, and furthermore, the literature does not provide any rigorous connections between the tail-index α and the choice of the stepsize, batch-size in SGD or the curvature of the objective at hand which is key to relate the tail-index to the generalization properties of SGD.

Before stating our theoretical results in detail, let us informally motivate our main method of analysis. Suppose the initial SGD iterate x_0 is in the domain of attraction[3] of a local minimum $x_\star$ of f which is smooth and well-approximated by a quadratic function in this basin. Under

[3] We say x_0 is in the domain of attraction of a local minimum $x_\star$, if gradient descent iterations to minimize f started at x_0 with sufficiently small stepsize converge to $x_\star$ as the number of iterations goes to infinity.

this assumption, by considering a first-order Taylor approximation of $\nabla f^{(i)}(x)$ around $x_\star$, we have

$$\nabla f^{(i)}(x) \approx \nabla f^{(i)}(x_\star) + \nabla^2 f^{(i)}(x_\star)(x - x_\star).$$

By using this approximation, we can approximate the SGD recursion (1.3) as:

$$\begin{aligned} x_k &\approx x_{k-1} - (\eta/b) \sum\nolimits_{i\in\Omega_k} \nabla^2 f^{(i)}(x_\star) x_{k-1} \\ &\quad + (\eta/b) \sum\nolimits_{i\in\Omega_k} \left(\nabla^2 f^{(i)}(x_\star) x_\star - \nabla f^{(i)}(x_\star) \right) \\ &=: (I - (\eta/b) H_k)\, x_{k-1} + q_k, \end{aligned} \tag{3.3}$$

where I denotes the identity matrix of appropriate size. Here, our main observation is that the SGD recursion can be approximated by an *affine stochastic recursion*. In this case, the map $\Psi_\Omega(x)$ is affine in x, and in addition to Theorem 1, we have access to the tools from Lyapunov stability theory (Srikant & Ying, 2019) and implicit renewal theory for investigating its statistical properties (Kesten, 1973; Goldie, 1991). In particular, Srikant & Ying (2019) study affine stochastic recursions subject to Markovian noise with a Lyapunov approach and show that the lower-order moments of the iterates can be made small as a function of the stepsize while they can be upper-bounded by the moments of a Gaussian random variable. In addition, they provide some examples where higher-order moments are infinite in steady-state. In the renewal theoretic approach, the object of interest would be the matrix $(I - \frac{\eta}{b} H_k)$ which determines the behavior of x_k: depending on the moments of this matrix, x_k can have heavy or light tails, or might even diverge.

In this study, we focus on the tail behavior of the SGD dynamics by analyzing it through the lens of implicit renewal theory. As, the recursion (3.3) is obtained by a quadratic approximation of the component functions $f^{(i)}$, which arises naturally in linear regression, we will consider a simplified setting and study it in great depth this dynamics in the case of linear regression. As opposed to prior work, this formalization will enable us to derive sharp characterizations of the tail-index and its dependency to the parameters η, b and the curvature as well as rate of convergence ρ to the stationary distribution. Our analysis technique lays the first steps for the analysis of more general objectives, and our experiments provide strong empirical support that our theory extends to deep learning settings.

We now focus on the case when f is a quadratic, which arises in linear regression:

$$\min\nolimits_{x\in\mathbb{R}^d} F(x) := (1/2)\mathbb{E}_{(a,y)\sim\mathcal{D}} \left[\left(a^T x - y \right)^2 \right], \tag{3.4}$$

where the data (a, y) comes from an unknown distribution $\mathcal{D}$ with support $\mathbb{R}^d \times \mathbb{R}$. Assume we have access to i.i.d. samples (a_i, y_i) from the distribution $\mathcal{D}$ where $\nabla f^{(i)}(x) = a_i(a_i^T x - y_i)$ is an unbiased estimator of the true gradient $\nabla F(x)$. The curvature, i.e. the value of second partial derivatives, of this objective around a minimum is determined by the Hessian matrix $\mathbb{E}(aa^T)$ which depends on the distribution of a. In this setting, SGD with batch-size b leads to the iterations

$$\begin{aligned} &x_k = M_k x_{k-1} + q_k \text{ with } M_k := I - (\eta/b) H_k, \\ &H_k := \sum\nolimits_{i\in\Omega_k} a_i a_i^T,\ q_k := (\eta/b) \sum\nolimits_{i\in\Omega_k} a_i y_i\,, \end{aligned} \tag{3.5}$$

where $\Omega_k := \{b(k-1)+1, b(k-1)+2, \dots, bk\}$ with $|\Omega_k| = b$. Here, for simplicity, we assume that we are in the one-pass regime (also called the streaming setting (Frostig et al., 2015; Jain et al., 2017; Gao et al., 2021)) where each sample is used only once without being recycled. Our purpose in this paper is to show that heavy tails can arise in SGD even in simple settings such as when the input data a_i is Gaussian, *without the necessity to have a heavy-tailed input data*[4]. Consequently, we make the following assumptions on the data throughout the paper:

(A1) a_i's are i.i.d. with a continuous distribution supported on $\mathbb{R}^d$ with all the moments finite. All the moments of a_i are finite.

(A2) y_i are i.i.d. with a continuous density whose support is $\mathbb{R}$ with all the moments finite.

We assume **(A1)** and **(A2)** throughout the paper, and they are satisfied in a large variety of cases, for instance when a_i and y_i are normally distributed. Let us introduce

$$h(s) := \lim\nolimits_{k\to\infty} \left(\mathbb{E} \| M_k M_{k-1} \dots M_1 \|^s \right)^{1/k}, \tag{3.6}$$

which arises in stochastic matrix recursions (see e.g. Buraczewski et al. (2014)) where $\|\cdot\|$ denotes the matrix 2-norm (i.e. largest singular value of a matrix). Since $\mathbb{E}\|M_k\|^s < \infty$ for all k and $s > 0$, we have $h(s) < \infty$. Let us also define $\Pi_k := M_k M_{k-1} \dots M_1$ and

$$\rho := \lim\nolimits_{k\to\infty} (2k)^{-1} \log \left(\text{largest eigenvalue of } \Pi_k^T \Pi_k \right). \tag{3.7}$$

The latter quantity is called the top Lyapunov exponent of the stochastic recursion (3.5). Furthermore, if ρ exists and is negative, it can be shown that a stationary distribution of the recursion (3.5) exists.

Note that by Assumption **(A1)**, the matrices $M_k = I - \frac{\eta}{b} H_k$ are i.i.d. and by Assumption **(A3)**, the Hessian matrix of the objective (3.4) satisfies $\mathbb{E}(aa^T) = \sigma^2 I_d$ where the

[4]Note that if the input data is heavy-tailed, the stationary distribution of SGD automatically becomes heavy-tailed; see Buraczewski et al. (2012) for details. In our context, the challenge is to identify the occurrence of the heavy tails when the distribution of the input data is light-tailed, such as a simple Gaussian distribution.

value of σ^2 determines the *curvature* around a minimum; smaller (larger) σ^2 implies the objective will grow slower (faster) around the minimum and the minimum will be flatter (sharper) (see e.g. Dinh et al. (2017)).

In the following, we show that the limit density has a polynomial tail with a tail-index given precisely by α, the unique critical value such that $h(\alpha) = 1$. The result builds on adapting the techniques developed in stochastic matrix recursions (Alsmeyer & Mentemeier, 2012; Buraczewski et al., 2016) to our setting. Our result shows that even in the simplest setting when the input data is i.i.d. without any heavy tail, SGD iterates can lead to a heavy-tailed stationary distribution with an infinite variance. To our knowledge, this is the first time such a phenomenon is proven in the linear regression setting.

Theorem 2. *Consider the SGD iterations (3.5). If $\rho < 0$ and there exists a unique positive α such that $h(\alpha) = 1$, then (3.5) admits a unique stationary solution x_∞ and the SGD iterations converge to x_∞ in distribution, where the distribution of x_∞ satisfies*

$$\lim_{t\to\infty} t^{\alpha}\mathbb{P}\left(u^T x_\infty > t\right) = e_\alpha(u)\,, \quad u \in \mathbb{S}^{d-1}\,, \quad (3.8)$$

for some positive and continuous function e_α on $\mathbb{S}^{d-1}$.

The proofs of Theorem 2 and all the following results in the main paper are given in the Supplementary Document. As Martin & Mahoney (2019); Şimşekli et al. (2020) provide numerical and theoretical evidence showing that the tail-index α of the density of the network weights is closely related to the generalization performance, where smaller α indicates better generalization, a natural question of practical importance is *how the tail-index depends on the parameters of the problem including the batch-size, dimension and the stepsize*. In order to have a more explicit characterization of the tail-index, we will make the following additional assumption for the rest of the paper which says that the input is Gaussian.

(A3) $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are Gaussian distributed for every i.

Under **(A3)**, next result shows that the formulas for ρ and $h(s)$ can be simplified. Let H be a matrix with the same distribution as H_k, and e_1 be the first basis vector. Define

$$\tilde{\rho} := \mathbb{E}\log\|(I - (\eta/b)H)\,e_1\|\,,$$
$$\tilde{h}(s) := \mathbb{E}\left[\|(I - (\eta/b)H)\,e_1\|^s\right] \text{ for } \rho < 0. \quad (3.9)$$

Theorem 3. *Assume **(A3)** holds. Consider the SGD iterations (3.5). If $\rho < 0$, then (i) there exists a unique positive α such that $h(\alpha) = 1$ and (3.8) holds; (ii) we have $\rho = \tilde{\rho}$ and $h(s) = \tilde{h}(s)$, where $\tilde{\rho}$ and $\tilde{h}(s)$ are defined in (3.9).*

This connection will allow us to get finer characterizations of the stepsize and batch-size choices that will provably lead to heavy tails with an infinite variance.

When input is not Gaussian (Theorem 2), the explicit formula (3.9) for ρ and $h(s)$ will not hold as an equality but it will become the following inequality:

$$\rho \le \hat{\rho} := \mathbb{E}\log\|(I - (\eta/b)H)\|\,,$$
$$h(s) \le \hat{h}(s) := \mathbb{E}\left[\|(I - (\eta/b)H)\|^s\right], \quad (3.10)$$

where ρ and $h(s)$ are defined by (3.6). This inequality is just a consequence of sub-multiplicativity of the norm of matrix products appearing in (3.6). If $\hat{\alpha}$ is such that $\hat{h}(\hat{\alpha}) = 1$, then by (3.10), $\hat{\alpha}$ is a lower bound on the tail-index α that satisfies $h(\alpha) = 1$ where h is defined as in (3.6). In other words, when the input is not Gaussian, we have $\hat{\alpha} \le \alpha$ and therefore $\hat{\alpha}$ serves as a lower bound on the tail-index. Finally, we remark that $\hat{\rho}$ and $\hat{h}(s)$ can help us check the conditions in Theorem 2. Since $\rho \le \hat{\rho}$, we have $\rho < 0$ when $\hat{\rho} < 0$. Moreover, $h(0) = \hat{h}(0) = 1$, and one can check that $h(s)$ is convex in s. When $\hat{\rho} < 0$, $\hat{h}'(0) = \hat{\rho} < 0$, and $h(s) \le \hat{h}(s) < 1$ for any sufficiently small $s > 0$. Under some mild assumption on the data distribution, one can check that $\liminf_{s\to\infty} h(s) > 1$ and thus there exists a unique positive α such that $h(\alpha) = 1$.

When input is Gaussian satisfying **(A3)**, due to the spherical symmetry of the Gaussian distribution, we have also $\rho = \hat{\rho}$, $h(s) = \hat{h}(s)$ (see Lemma (16)). Furthermore, in this case, by using the explicit characterization of the tail-index α in Theorem 3, we prove that larger batch-sizes lead to a lighter tail (i.e. larger α), which links the heavy tails to the observation that smaller b yields improved generalization in a variety of settings in deep learning (Keskar et al., 2017; Panigrahi et al., 2019; Martin & Mahoney, 2019). We also prove that smaller stepsizes lead to larger α, hence lighter tails, which agrees with the fact that the existing literature for linear regression often choose η small enough to guarantee that variance of the iterates stay bounded (Dieuleveut et al., 2017; Jain et al., 2017).

Theorem 4. *Assume **(A3)** holds. The tail-index α is strictly increasing in batch-size b and strictly decreasing in stepsize η and variance σ^2 provided that $\alpha \ge 1$. Moreover, the tail-index α is strictly decreasing in dimension d.*

When input is not Gaussian, Theorem 4 can be adapted in the sense that $\hat{\alpha}$ (defined via $\hat{h}(\hat{\alpha}) = 1$) will be strictly increasing in batch-size b and strictly increasing in stepsize η and variance σ^2 provided that $\hat{\alpha} \ge 1$.

Under **(A3)**, next result characterizes the tail-index α depending on the choice of the batch-size b, the variance σ^2, which determines the curvature around the minimum and the stepsize; in particular we show that if the stepsize exceeds an explicit threshold, the stationary distribution will become heavy tailed with an infinite variance.

Proposition 5. *Assume **(A3)** holds. Let $\eta_{crit} = \frac{2b}{\sigma^2(d+b+1)}$. The following holds: (i) There exists $\eta_{max} > \eta_{crit}$ such that*

for any $\eta_{crit} < \eta < \eta_{max}$, Theorem 2 holds with tail-index $0 < \alpha < 2$. (ii) If $\eta = \eta_{crit}$, Theorem 2 holds with tail-index $\alpha = 2$. (iii) If $\eta \in (0, \eta_{crit})$, then Theorem 2 holds with tail-index $\alpha > 2$.

Relation to first exit times. Proposition 5 implies that, for fixed η and b, the tail-index α will be decreasing with increasing σ. Combined with the first-exit-time analyses of Şimşekli et al. (2019b); Nguyen et al. (2019), which state that the escape probability from a basin becomes higher for smaller α, our result implies that the probability of SGD escaping from a basin gets larger with increasing curvature; hence providing an alternative view for the argument that SGD prefers flat minima.

Three regimes for stepsize. Theorems 2-4 and Proposition 5 identify three regimes: (I) convergence to a limit with a finite variance if $\rho < 0$ and $\alpha > 2$; (II) convergence to a heavy-tailed limit with infinite variance if $\rho < 0$ and $\alpha < 2$; (III) $\rho > 0$ when convergence cannot be guaranteed. For Gaussian input, if the stepsize is small enough, smaller than η_{crit}, by Proposition 5, $\rho < 0$ and $\alpha > 2$, therefore regime (I) applies. As we increase the stepsize, there is a critical stepsize level η_{crit} for which $\eta > \eta_{crit}$ leads to $\alpha < 2$ as long as $\eta < \eta_{max}$ where η_{max} is the maximum allowed stepsize for ensuring convergence (corresponds to $\rho = 0$). A similar behavior with three (learning rate) stepsize regimes was reported in Lewkowycz et al. (2020) and derived analytically for one hidden layer linear networks with a large width. The large stepsize choices that avoids divergence, so called the *catapult phase* for the stepsize, yielded the best generalization performance empirically, driving the iterates to a flatter minima in practice. We suspect that the catapult phase in Lewkowycz et al. (2020) corresponds to regime (II) in our case, where the iterates are heavy-tailed, which might cause convergence to flatter minima as the first-exit-time discussions suggest (Şimşekli et al., 2019a).

Moment Bounds and Convergence Speed. Theorem 2 is of asymptotic nature which characterizes the stationary distribution x_∞ of SGD iterations with a tail-index α. Next, we provide non-asymptotic moment bounds for x_k at each k-th iterate, and also for the limit x_∞.

Theorem 6. *Assume **(A3)** holds. (i) If the tail-index $\alpha \leq 1$, then for any $p \in (0, \alpha)$, we have $h(p) < 1$ and $\mathbb{E}\|x_k\|^p \leq (h(p))^k \mathbb{E}\|x_0\|^p + \frac{1-(h(p))^k}{1-h(p)} \mathbb{E}\|q_1\|^p$. (ii) If the tail-index $\alpha > 1$, then for any $p \in (1, \alpha)$, we have $h(p) < 1$ and for any $0 < \epsilon < \frac{1}{h(p)} - 1$, we have $\mathbb{E}\|x_k\|^p \leq ((1+\epsilon)h(p))^k \mathbb{E}\|x_0\|^p + \frac{1-((1+\epsilon)h(p))^k}{1-(1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{((1+\epsilon)^{\frac{1}{p-1}} - 1)^p} \mathbb{E}\|q_1\|^p$.*

Theorem 6 shows that when $p < \alpha$ the upper bound on the p-th moment of the iterates converges exponentially to the p-the moment of q_1 when $\alpha \leq 1$ and a neighborhood of the p-moment of q_1 when $\alpha > 1$, where q_1 is defined in (3.5).

By letting $k \to \infty$ and applying Fatou's lemma, we can also characterize the moments of the stationary distribution.

Corollary 7. *Assume **(A3)** holds. (i) If $\alpha \leq 1$, then for any $p \in (0, \alpha)$, $\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-h(p)} \mathbb{E}\|q_1\|^p$, where $h(p) < 1$. (ii) If $\alpha > 1$, then for any $p \in (1, \alpha)$, we have $h(p) < 1$ and for any $\epsilon > 0$ such that $(1+\epsilon)h(p) < 1$, we have $\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-(1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{((1+\epsilon)^{\frac{1}{p-1}} - 1)^p} \mathbb{E}\|q_1\|^p$.*

Next, we will study the speed of convergence of the k-th iterate x_k to its stationary distribution x_∞ in the Wasserstein metric $\mathcal{W}_p$ for any $1 \leq p < \alpha$.

Theorem 8. *Assume **(A3)** holds. Assume $\alpha > 1$. Let ν_k, ν_∞ denote the probability laws of x_k and x_∞ respectively. Then $\mathcal{W}_p(\nu_k, \nu_\infty) \leq (h(p))^{k/p} \mathcal{W}_p(\nu_0, \nu_\infty)$, for any $1 \leq p < \alpha$, where the convergence rate $(h(p))^{1/p} \in (0, 1)$.*

Theorem 8 shows that in case $\alpha < 2$ the convergence to a heavy tailed distribution occurs relatively fast, i.e. with a linear convergence in the p-Wasserstein metric. We can also characterize the constant $h(p)$ in Theorem 8 which controls the convergence rate as follows:

Corollary 9. *Assume **(A3)** holds. When $\eta < \eta_{crit} = \frac{2b}{\sigma^2(d+b+1)}$, we have the tail-index $\alpha > 2$, and*

$$\mathcal{W}_2(\nu_k, \nu_\infty) \leq \left(1 - 2\eta\sigma^2 \left(1 - \eta/\eta_{crit}\right)\right)^{k/2} \mathcal{W}_2(\nu_0, \nu_\infty).$$

Theorem 8 works for any $p < \alpha$. At the critical $p = \alpha$, Theorem 2 indicates that $\mathbb{E}\|x_\infty\|^\alpha = \infty$, and therefore we have $\mathbb{E}\|x_k\|^\alpha \to \infty$ as $k \to \infty$, [5] which serves as an evidence that the tail gets heavier as the number of iterates k increases. By adapting the proof of Theorem 6, we have the following result stating that the moments of the iterates of order α go to infinity but this speed can only be polynomially fast.

Proposition 10. *Assume **(A3)** holds. Given the tail-index α, we have $\mathbb{E}\|x_\infty\|^\alpha = \infty$. Moreover, $\mathbb{E}\|x_k\|^\alpha = O(k)$ if $\alpha \leq 1$, and $\mathbb{E}\|x_k\|^\alpha = O(k^\alpha)$ if $\alpha > 1$.*

It may be possible to leverage recent results on the concentration of products of i.i.d. random matrices (Huang et al., 2020; Henriksen & Ward, 2020) to study the tail of x_k for finite k, which can be a future research direction.

Generalized Central Limit Theorem for Ergodic Averages. When $\alpha > 2$, by Corollary 7, second moment of the iterates x_k are finite, in which case central limit theorem (CLT) says that if the cumulative sum of the iterates $S_K = \sum_{k=1}^K x_k$ is scaled properly, the resulting distribution is Gaussian. In the case where $\alpha < 2$, the variance of the iterates is not finite; however in this case, we derive the following generalized CLT (GCLT) which says if the iterates

[5] Otherwise, one can construct a subsequence x_{n_k} that is bounded in the space L^α converging to x_∞ which would be a contradiction.

are properly scaled, the limit will be an α-stable distribution. This is stated in a more precise manner as follows.

Corollary 11. *Assume the conditions of Theorem 2 are satisfied, i.e. assume $\rho < 0$ and there exists a unique positive α such that $h(\alpha) = 1$. Then, we have the following:*

(i) If $\alpha \in (0,1) \cup (1,2)$, then there is a sequence $d_K = d_K(\alpha)$ and a function $C_\alpha : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-\frac{1}{\alpha}}(S_K - d_K)$ converge in law to the α-stable random variable with characteristic function $\Upsilon_\alpha(tv) = \exp(t^\alpha C_\alpha(v))$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(ii) If $\alpha = 1$, then there are functions $\xi, \tau : (0,\infty) \mapsto \mathbb{R}$ and $C_1 : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-1}S_K - K\xi\left(K^{-1}\right)$ converge in law to the random variable with characteristic function $\Upsilon_1(tv) = \exp\left(tC_1(v) + it\langle v, \tau(t)\rangle\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iii) If $\alpha = 2$, then there is a sequence $d_K = d_K(2)$ and a function $C_2 : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ such that as $K \to \infty$ the random variables $(K \log K)^{-\frac{1}{2}}(S_K - d_K)$ converge in law to the random variable with characteristic function $\Upsilon_2(tv) = \exp\left(t^2 C_2(v)\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iv) If $\alpha \in (0,1)$, then $d_K = 0$, and if $\alpha \in (1,2]$, then $d_K = K\bar{x}$, where $\bar{x} = \int_{\mathbb{R}^d} x\nu_\infty(dx)$.

In addition to its evident theoretical interest, Corollary 11 has also an important practical implication: estimating the tail-index of a *generic* heavy-tailed distribution is a challenging problem (see e.g. Clauset et al. (2009); Goldstein et al. (2004); Bauke (2007)); however, for the specific case of α-stable distributions, accurate and computationally efficient estimators, which *do not* require the knowledge of the functions C_α, τ, ξ, have been proposed (Mohammadi et al., 2015). Thanks to Corollary 11, we will be able to use such estimators in our numerical experiments in Section 4.

Further Discussions. Even though we focus on the case when f is quadratic and consider the linear regression (3.4), for fully non-convex Lipschitz losses, it is possible to show that a stationary distribution exists and the distribution of the iterates converge to it exponentially fast but heavy-tails in this setting is not understood in general except some very special cases; see e.g. Diaconis & Freedman (1999). However, if the gradients have asymptotic linear growth, even for non-convex objectives, extending our tail index results beyond quadratic optimization is possible if we incorporate the proof techniques of Alsmeyer (2016) to our setting. However, in this more general case, characterizing the tail index explicitly and studying its dependence on stepsize, batch-size does not seem to be a tractable problem since the dependence of the asymptotic linear growth of the random iteration on the data may not be tractable, therefore studying the quadratic case allows us a deeper understanding of the tail index on a relatively simpler problem.

We finally note that the gradient noise in SGD is actually both multiplicative and additive (Dieuleveut et al., 2017; 2020); a fact that is often discarded for simplifying the mathematical analysis. In the linear regression setting, we have shown that the multiplicative noise M_k is the main source of heavy-tails, where a deterministic M_k would not lead to heavy tails.[6] In light of our theory, in Section A in the Supplementary Document, we discuss in detail the recently proposed stochastic differential equation (SDE) representations of SGD in continuous-time and argue that, compared to classical SDEs driven by a Brownian motion (Jastrzębski et al., 2017; Cheng et al., 2020), SDEs driven by heavy-tailed α-stable Lévy processes (Şimşekli et al., 2019b) are more adequate when $\alpha < 2$.

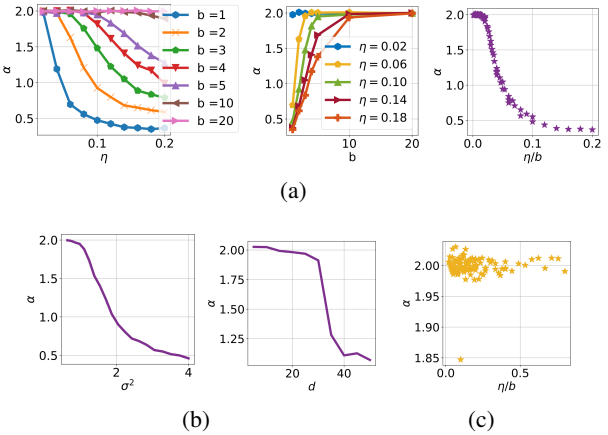
4. Experiments

In this section, we present our experimental results on both synthetic and real data, in order to illustrate that our theory also holds in finite-sum problems (besides the streaming setting). Our main goal will be to illustrate the tail behavior of SGD by varying the algorithm parameters: depending on the choice of the stepsize η and the batch-size b, the distribution of the iterates does converge to a heavy-tailed distribution (Theorem 2) and the behavior of the tail-index obeys Theorem 4. Our implementations can be found in `github.com/umutsimsekli/sgd_ht`.

Synthetic experiments. In our first setting, we consider a simple synthetical setup, where we assume that the data points follow a Gaussian distribution. We will illustrate that the SGD iterates can become heavy-tailed even in this simplistic setting where the problem is a simple linear regression with all the variables being Gaussian. More precisely, we will consider the following model: $x_0 \sim \mathcal{N}(0, \sigma_x^2 I)$, $a_i \sim \mathcal{N}(0, \sigma^2 I)$, and $y_i | a_i, x_0 \sim \mathcal{N}\left(a_i^\top x_0, \sigma_y^2\right)$, where $x_0, a_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ for $i = 1, \dots, n$, and $\sigma, \sigma_x, \sigma_y > 0$.

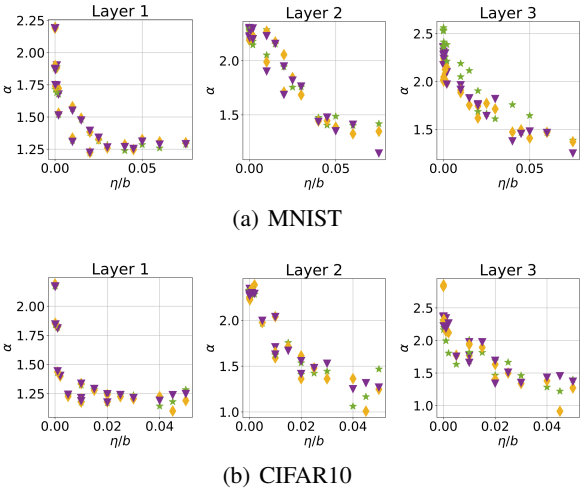
In our experiments, we will need to estimate the tail-index α of the stationary distribution ν_∞. Even though several tail-index estimators have been proposed for generic heavy-tailed distributions in the literature (Paulauskas & Vaičiulis, 2011), we observed that, even for small d, these estimators can yield inaccurate estimations and require tuning hyperparameters, which is non-trivial. We circumvent this issue thanks to the GCLT in Corollary 11: since the average of the iterates is guaranteed to converge to a multivariate α-stable random variable in distribution, we can use the tail-index estimators that are specifically designed for stable distributions. By following Tzagkarakis et al. (2018); Şimşekli et al. (2019b), we use the estimator proposed by Mohammadi et al. (2015), which is fortunately agnostic to the scaling function

[6] E.g., if M_k is deterministic and q_k is Gaussian, then x_k is Gaussian for all k, and so is x_∞ if the limit exists.



(a) (b) (c)

Figure 1. Behavior of α with (a) varying stepsize η and batch-size b, (b) d and σ, (c) under RMSProp.



(a) MNIST

(b) CIFAR10

Figure 2. Results on FCNs. Different markers represent different initializations with the same η, b.

C_α. The details of this estimator are given in Section B in the Supplementary Document.

To benefit from the GCLT, we are required to compute the average of the 'centered' iterates: $\frac{1}{K-K_0}\sum_{k=K-K_0+1}^{K}(x_k - \bar{x})$, where K_0 is a 'burn-in' period aiming to discard the initial phase of SGD, and the mean of ν_∞ is given by $\bar{x} = \int_{\mathbb{R}^d} x\nu_\infty(dx) = (A^\top A)^{-1}A^\top y$ as long as $\alpha > 1$[7], where the i-th row of $A \in \mathbb{R}^{n\times d}$ contains $a_i^\top$ and $y = [y_1, \dots, y_n] \in \mathbb{R}^n$. We then repeat this procedure 1600 times for different initial points and obtain 1600 different random vectors, whose distributions are supposedly close to an α-stable distribution. Finally, we run the tail-index estimator of Mohammadi et al. (2015) on these random vectors to estimate α.

In our first experiment, we investigate the tail-index α of the stationary measure ν_∞ for varying stepsize η and batch-size b. We set $d = 100$ first fix the variances $\sigma = 1$, $\sigma_x = \sigma_y = 3$, and generate $\{a_i, y_i\}_{i=1}^n$ by simulating the statistical model. Then, by fixing this dataset, we run the SGD recursion (3.5) for a large number of iterations and vary η from 0.02 to 0.2 and b from 1 to 20. We also set $K = 1000$ and $K_0 = 500$. Figure 1(a) illustrates the results. We can observe that, increasing η and decreasing b both result in decreasing α, where the tail-index can be prohibitively small (i.e., $\alpha < 1$, hence even the mean of ν_∞ is not defined) for large η. Besides, we can also observe that the tail-index is in strong correlation with the ratio η/b.

In our second experiment, we investigate the effect of d and σ on α. In Figure 1(b) (left), we set $d = 100$, $\eta = 0.1$ and $b = 5$ and vary σ from 0.8 to 2. For each value of σ, we simulate a new dataset from by using the generative model and run SGD with K, K_0. We again repeat each experiment 1600 times. We follow a similar route for Figure 1(b) (right): we fix $\sigma = 1.75$ and repeat the previous procedure for each value of d ranging from 5 to 50. The results confirm our theory: α decreases for increasing σ and d, and we observe that for a fixed b and η the change in d can abruptly alter α.

In our final synthetic data experiment, we investigate how the tails behave under adaptive optimization algorithms. We replicate the setting of our first experiment, with the only difference that we replace SGD with RMSProp (Hinton et al., 2012). As shown in Figure 1(c), the 'clipping' effect of RMSProp as reported in Zhang et al. (2020); Zhou et al. (2020) prevents the iterates become heavy-tailed and the vast majority of the estimated tail-indices is around 2, indicating a Gaussian behavior. On the other hand, we repeated the same experiment with the variance-reduced optimization algorithm SVRG (Johnson & Zhang, 2013), and observed that for almost all choices of η and b the algorithm converges near the minimizer (with an error in the order of 10^{-6}), hence the stationary distribution ν_∞ seems to be a degenerate distribution, which does not admit a heavy-tailed behavior. Regarding the link between heavy-tails and generalization (Martin & Mahoney, 2019; Şimşekli et al., 2020), this behavior of RMSProp and SVRG might be related to their ineffective generalization as reported in Keskar & Socher (2017); Defazio & Bottou (2019).

Experiments on fully connected neural networks. In the second set of experiments, we investigate the applicability of our theory beyond the quadratic optimization problems. Here, we follow the setup of Şimşekli et al. (2019a) and consider a fully connected neural network with the cross entropy loss and ReLU activation functions on the MNIST and CIFAR10 datasets. We train the models by using SGD

[7] The form of $\bar{x}$ can be verified by noticing that $\mathbb{E}[x_k]$ converges to the minimizer of the problem by the law of total expectation. Besides, our GCLT requires the sum of the iterates to be normalized by $\frac{1}{(K-K_0)^{1/\alpha}}$; however, for a finite K, normalizing by $\frac{1}{K-K_0}$ results in a scale difference, to which our estimator is agnostic.

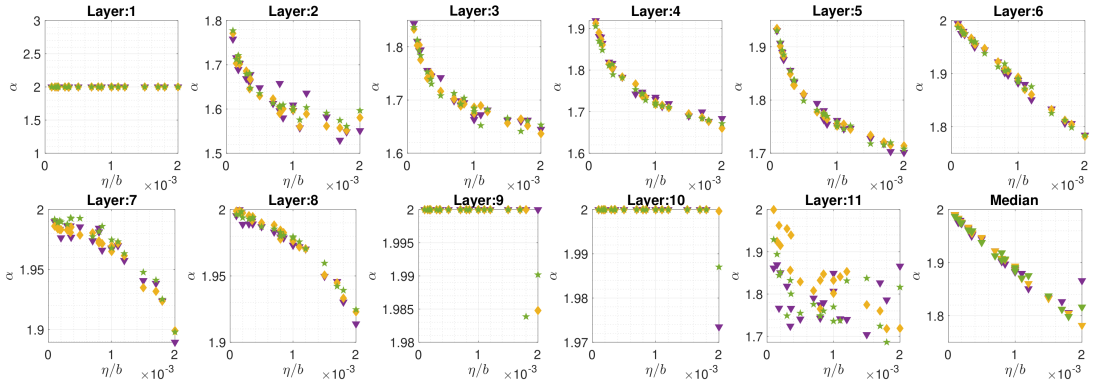


Figure 3. Results on VGG networks. The values of α that exceeded 2 is truncated to 2 for visualization purposes. Different markers represent different initializations.

for 10K iterations and we range η from 10^{-4} to 10^{-1} and b from 1 to 10. Since it would be computationally infeasible to repeat each run thousands of times as we did in the synthetic data experiments, in this setting we follow a different approach based on (i) (Şimşekli et al., 2019a) that suggests that the tail behavior can differ in different layers of a neural network, and (ii) (De Bortoli et al., 2020) that shows that in the infinite width limit, the different components of a given layer of a two-layer fully connected network (FCN) becomes independent. Accordingly, we first compute the average of the last 1K SGD iterates, whose distribution should be close an α-stable distribution by the GCLT. We then treat each layer as a collection of i.i.d. α-stable random variables and measure the tail-index of each individual layer separately by using the the estimator from Mohammadi et al. (2015). Figure 2 shows the results for a three-layer network (with 128 hidden units at each layer) , whereas we obtained very similar results with a two-layer network as well. We observe that, while the dependence of α on η/b differs from layer to layer, in each layer the measured α correlate very-well with the ratio η/b in both datasets.

Experiments on VGG networks. In our last set of experiments, we evaluate our theory on VGG networks (Simonyan & Zisserman, 2015) on CIFAR10 with 11 layers (10 convolutional layers with max-pooling and ReLU units, followed by a final linear layer), which contains 10M parameters. We follow the same procedure as we used for the fully connected networks, where we vary η from 10^{-4} to 1.7×10^{-3} and b from 1 to 10. The results are shown in Figure 3. Similar to the previous experiments, we observe that α depends on the layers. For the layers 2-8, the tail-index correlates well with the ratio η/b, whereas the first and layers 1, 9, and 10 exhibit a Gaussian behavior ($\alpha \approx 2$). On the other hand, the correlation between the tail-index of the last layer (which is linear) with η/b is still visible, yet less clear. Finally, in the last plot, we compute the median of the estimate tail-indices over layers, and observe a very clear decrease with increasing η/b. These observations provide further support for our theory and show that the heavy-tail phenomenon also occurs in neural networks, whereas α is potentially related to η and b in a more complicated way.

5. Conclusion and Future Directions

We studied the tail behavior of SGD and showed that depending on η, b and the curvature, the iterates can converge to a *heavy-tailed* random variable in distribution. We further supported our theory with various experiments conducted on neural networks and illustrated that our results would also apply to more general settings and hence provide new insights about the behavior of SGD in deep learning. Our study also brings up a number of future directions. (i) Our proof techniques are for the streaming setting, where each sample is used only once. Extending our results to the finite-sum scenario and investigating the effects of finite-sample size on the tail-index would be an interesting future research direction. (ii) We suspect that the tail-index may have an impact on the time required to escape a saddle point and this can be investigated further as another future research direction. (iii) Our work considers SGD with constant stepsize. Extending our analysis to adaptive methods and varying stepsizes is another interesting future research direction.

Acknowledgements. M.G.'s research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. U.Ş.'s research is supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). L.Z. is grateful to the support from a Simons Foundation Collaboration Grant and the grant NSF DMS-2053454 from the National Science Foundation.

References

Ali, A., Dobriban, E., and Tibshirani, R. J. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 233–244, 2020.

Alsmeyer, G. On the stationary tail index of iterated random Lipschitz functions. *Stochastic Processes and their Applications*, 126(1):209–233, 2016.

Alsmeyer, G. and Mentemeier, S. Tail behaviour of stationary solutions of random difference equations: the case of regular matrices. *Journal of Difference Equations and Applications*, 18(8):1305–1332, 2012.

Bauke, H. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B*, 58(2):167–173, 2007.

Ben Arous, G. and Guionnet, A. The spectrum of heavy tailed random matrices. *Communications in Mathematical Physics*, 278(3):715–751, 2008.

Bertoin, J. *Lévy Processes*. Cambridge University Press, 1996.

Buraczewski, D., Damek, E., and Mirek, M. Asymptotics of stationary solutions of multivariate stochastic recursions with heavy tailed inputs and related limit theorems. *Stochastic Processes and their Applications*, 122(1):42–67, 2012.

Buraczewski, D., Damek, E., Guivarc'h, Y., and Mentemeier, S. On multidimensional Mandelbrot cascades. *Journal of Difference Equations and Applications*, 20 (11):1523–1567, 2014.

Buraczewski, D., Damek, E., and Przebinda, T. On the rate of convergence in the Kesten renewal theorem. *Electronic Journal of Probaiblity*, 20(22):1–35, 2015.

Buraczewski, D., Damek, E., and Mikosch, T. *Stochastic Models with Power-Law Tails*. Springer, 2016.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Cheng, X., Yin, D., Bartlett, P. L., and Jordan, M. I. Stochastic gradient and Langevin processes. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1810–1819, 2020.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM Review*, 51(4): 661–703, 2009.

De Bortoli, V., Durmus, A., Fontaine, X., and Şimşekli, U. Quantitative propagation of chaos for SGD in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Defazio, A. and Bottou, L. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pp. 1755–1765, 2019.

Diaconis, P. and Freedman, D. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1019–1028. JMLR. org, 2017.

Fink, H. and Klüppelberg, C. Fractional Lévy-driven Ornstein–Uhlenbeck processes and stochastic differential equations. *Bernoulli*, 17(1):484–506, 2011.

Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, pp. 728–763, 2015.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Global convergence of stochastic gradient hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *To Appear, Operations Research*, 2021.

Goldie, C. M. Implicit renewal theory and tails of solutions of random equations. *Annals of Applied Probability*, 1 (1):126–166, 1991.

Goldstein, M. L., Morris, S. A., and Yen, G. G. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

Henriksen, A. and Ward, R. Concentration inequalities for random matrix products. *Linear Algebra and its Applications*, 594:81–94, 2020.

Hinton, G., Srivastava, N., and Swersky, K. Overview of mini-batch gradient descent. *Neural Networks for Machine Learning*, Lecture 6a, 2012. URL

http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Hodgkinson, L. and Mahoney, M. W. Multiplicative noise and heavy tails in stochastic optimization. *arXiv preprint arXiv:2006.06293*, June 2020.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Science and Applications*, 4(1): 3–32, 2019.

Huang, D., Niles-Weed, J., Tropp, J. A., and Ward, R. Matrix concentration for products. *arXiv preprint arXiv:2003.05437*, 2020.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. Accelerating stochastic gradient descent. In *Proc. STAT*, volume 1050, pp. 26, 2017.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Keskar, N. S. and Socher, R. Improving generalization performance by switching from Adam to SGD. *arXiv preprint arXiv:1712.07628*, 2017.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Kesten, H. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131: 207–248, 1973.

Lévy, P. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1937.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2101–2110, 06–11 Aug 2017.

Mandt, S., Hoffman, M. D., and Blei, D. M. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 354–363, 2016.

Martin, C. H. and Mahoney, M. W. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Mirek, M. Heavy tail phenomenon and convergence to stable laws for iterated Lipschitz maps. *Probability Theory and Related Fields*, 151(3-4):705–734, 2011.

Mohammadi, M., Mohammadpour, A., and Ogata, H. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

Newman, C. M. The distribution of Lyapunov exponents: Exact results for random matrices. *Communications in Mathematical Physics*, 103(1):121–126, 1986.

Nguyen, T. H., Şimşekli, U., Gürbüzbalaban, M., and Richard, G. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, pp. 273–283, 2019.

Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2013.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-Gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

Paulauskas, V. and Vaičiulis, M. Once more on comparison of tail index estimators. *arXiv preprint arXiv:1104.1242*, 2011.

Pavlyukevich, I. Cooling down Lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 40(41): 12299–12313, 2007.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019a.

Şimşekli, U., Sagun, L., and Gürbüzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837, 2019b.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Tzagkarakis, G., Nolan, J. P., and Tsakalides, P. Compressive sensing of temporally correlated sources using isotropic multivariate stable laws. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1710–1714. IEEE, 2018.

Villani, C. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Xie, Z., Sato, I., and Sugiyama, M. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S., and E, W. Towards theoretically understanding why SGD generalizes better than ADAM in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.