arXiv:1901.06053v1 [cs.LG] 18 Jan 2019

A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks

Umut Şimşekli*, Levent Sagun†, Mert Gürbüzbalaban ‡

Abstract

The gradient noise (GN) in the stochastic gradient descent (SGD) algorithm is often considered to be Gaussian in the large data regime by assuming that the *classical* central limit theorem (CLT) kicks in. This assumption is often made for mathematical convenience, since it enables SGD to be analyzed as a stochastic differential equation (SDE) driven by a Brownian motion. We argue that the Gaussianity assumption might fail to hold in deep learning settings and hence render the Brownian motion-based analyses inappropriate. Inspired by non-Gaussian natural phenomena, we consider the GN in a more general context and invoke the *generalized* CLT (GCLT), which suggests that the GN converges to a *heavy-tailed* α-stable random variable. Accordingly, we propose to analyze SGD as an SDE driven by a Lévy motion. Such SDEs can incur 'jumps', which force the SDE *transition* from narrow minima to wider minima, as proven by existing metastability theory. To validate the α-stable assumption, we conduct extensive experiments on common deep learning architectures and show that in all settings, the GN is highly non-Gaussian and admits heavy-tails. We further investigate the tail behavior in varying network architectures and sizes, loss functions, and datasets. Our results open up a different perspective and shed more light on the belief that SGD prefers wide minima.

1 Introduction

Context and motivation: Deep neural networks have revolutionized machine learning and have ubiquitous use in many application domains [19, 28, 31]. In full generality, many key tasks in deep learning reduces to solving the following optimization problem:

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \Big\{ f(\mathbf{w}) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} f^{(i)}(\mathbf{w}) \Big\} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^p$ denotes the weights of the neural network, $f : \mathbb{R}^p \to \mathbb{R}$ denotes the loss function that is typically non-convex in $\mathbf{w}$, each $f^{(i)}$ denotes the (instantaneous) loss function that is contributed by the *data point* $i \in \{1, \dots, n\}$, and n denotes the total number of data points. Stochastic gradient descent (SGD) is one the most popular approaches for attacking this problem in practice and is based on the following iterative updates:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla \tilde{f}_k(\mathbf{w}_k) \tag{2}$$

where $k \in \{1, \dots, K\}$ denotes the iteration number and $\nabla \tilde{f}_k$ denotes the stochastic gradient at iteration k, that is defined as follows:

$$\nabla \tilde{f}_k(\mathbf{w}) \triangleq \nabla \tilde{f}_{\Omega_k}(\mathbf{w}) \triangleq \frac{1}{b}\sum\nolimits_{i\in\Omega_k} \nabla f^{(i)}(\mathbf{w}). \tag{3}$$

Here, $\Omega_k \subset \{1, \dots, n\}$ is a random subset that is drawn with or without replacement at iteration k, and $b = |\Omega_k|$ denotes the number of elements in Ω_k.

SGD is widely used in deep learning with a great success in its computational efficiency [4, 5]. Beyond efficiency, understanding how SGD performs better than its full batch counterpart in terms of test accuracy remains a major challenge. Even though SGD seems to find zero loss solutions on

*LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France.
†Institute of Physics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.
‡Department of Management Science and Information Systems, Rutgers Business School, NJ 08854, USA.

the training landscape (at least in certain regimes [17, 27, 47, 59]), it appears that the algorithm finds solutions with different properties depending on how it is tuned [21, 25, 27, 37, 49, 50]. Despite the fact that the impact of SGD on generalization has been studied [1, 40, 54], a satisfactory theory that can explain its success in a way that encompasses such peculiar empirical properties is still lacking.

A popular approach for analyzing SGD is based on considering SGD as a discretization of a continuous-time process [9, 22, 25, 33, 36, 61]. This approach mainly requires the following assumption[1] on the stochastic gradient noise $U_k(\mathbf{w}) \triangleq \nabla \tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})$:

$$U_k(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4}$$

where $\mathcal{N}$ denotes the multivariate (Gaussian) normal distribution and $\mathbf{I}$ denotes the identity matrix of appropriate size. The rationale behind this assumption is that, if the size of the minibatch b is large enough, then we can invoke the Central Limit Theorem (CLT) and assume that the distribution of U_k is approximately Gaussian. Then, under this assumption, (2) can be written as follows:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k) + \sqrt{\eta}\sqrt{\eta\sigma^2} Z_k, \tag{5}$$

where Z_k denotes a standard normal random variable in $\mathbb{R}^p$. If we further assume that the step-size η is small enough, then the continuous-time analogue of the discrete-time process (5) is the following stochastic differential equation (SDE):[2]

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \sqrt{\eta\sigma^2}\mathrm{dB}_t, \tag{6}$$

where B_t denotes the standard Brownian motion. This SDE is a variant of the well-known *Langevin diffusion* and under mild regularity assumptions on f, one can show that the Markov process $(\mathbf{w}_t)_{t\geq 0}$ is ergodic with its unique invariant measure, whose density is proportional to $\exp(-f(x)/(\eta\sigma^2))$ for any $\eta > 0$. [45]. From this perspective, the SGD recursion in (5) can be seen as a first-order Euler-Maruyama discretization of the Langevin dynamics (see also [22, 25, 33]), which is often referred to as the Unadjusted Langevin Algorithm (ULA) [14, 30, 45].

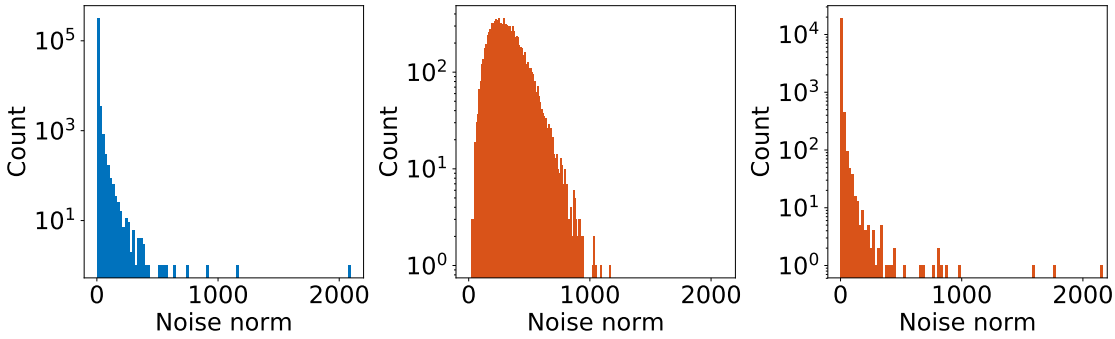
Based on this observation, [25] focused on the relation between this invariant measure and the algorithm parameters, namely the step-size η and mini-batch size, as a function of σ^2. They concluded that the ratio of learning rate divided by the batch size is the control parameter that determines the width of the minima found by SGD. Furthermore, they revisit the famous wide minima folklore [20]: Among the minima found by SGD, the wider it is, the better it performs on the test set. However, there are several fundamental issues with this approach, which we will explain below.

We first illustrate a typical mismatch between the Gaussianity assumption and the empirical behavior of the stochastic gradient noise. In Figure 1, we plot the histogram of the norms of the stochastic gradient noise that is computed using a convolutional neural network in a real classification problem and compare it to the histogram of the norms of Gaussian random variables. It can be clearly observed that the shape of the real histogram is very different than the Gaussian and shows a *heavy-tailed* behavior.

In addition to the empirical observations, the Gaussianity assumption also yields some theoretical issues. The first issue with this assumption is that the current SDE analyses of SGD are based on the *invariant measure* of the SDE, which implicitly assumes that sufficiently many iterations have been taken to converge to that measure. Recent results on ULA [44, 56] have shown that, the required number of iterations to achieve the invariant measure often grows exponentially with the dimension p. This result contradicts with the current practice: considering the large size of the neural networks and limited computational budget, only a limited number of iterations – which is much smaller than $\exp(\mathcal{O}(p))$ – can be taken. This conflict becomes clearer in the light of the recent works that studied the *local* behavior of ULA [51, 60]. These studies showed that ULA will get close to the nearest local optimum in polynomial time; however, the required amount of

[1] We note that more sophisticated assumptions than (4) have been made in terms of the covariance matrix of the Gaussian distribution (e.g. state dependent, anisotropic). However, in all these cases, the resulting distribution is still a Gaussian, therefore the same criticism holds.

[2] In a recent work with a similar critic taken on the recent theories on the SGD dynamics, some theoretical concerns have been also raised about the SDE approximation of SGD [57]. We believe that the SDE representation is sufficiently accurate for small step-sizes and a good, if not the best, proxy for understanding the behavior of SGD.



(a) Real (b) Gaussian (c) α-stable

Figure 1: (a)The histogram of the norm of the gradient noises computed with AlexNet on Cifar10. (b) and (c) the histograms of the norms of (scaled) Guassian and α-stable random variables.

time for escaping from that local optimum increases exponentially with the dimension. Therefore, the phenomenon that SGD prefers wide minima within a considerably small number of iterations cannot be explained using the asymptotic distribution of the SDE given in (6).

The second issue is related to the local behavior of the process and becomes clear when we consider the *metastability* analysis of Brownian motion-driven SDEs. These studies [6, 16, 24] consider the case where $\mathbf{w}_0$ is initialized in a quadratic basin and then analyze the minimum time t such that $\mathbf{w}_t$ is outside that basin. They show that this so-called *first exit time* depends *exponentially* on the height of the basin; however, this dependency is only *polynomial* with the width of the basin. These theoretical results directly contradict with the the wide minima phenomenon: even if the height of a basin is slightly larger, the exit-time from this basin will be dominated by its height, which implies that the process would stay longer in (or in other words, 'prefer') deeper minima as opposed to wider minima. The reason why the exit-time is dominated by the height is due to the *continuity* of the Brownian motion, which is in fact a direct consequence of the Gaussian noise assumption.

A final remark on the issues of this approach is the observation that landscape is flat at the bottom regardless of the batch size used in SGD [46]. In particular, the spectrum of the Hessian at a near critical point with close to zero loss value has many near zero eigenvalues. Therefore, local curvature measures that are used as a proxy for measuring the width of a basin correlates with the magnitudes of large eigenvalues of the Hessian which are few. Besides, during the dynamics of SGD it has been observed that the algorithm does not cross barriers except perhaps at the very initial phase [2, 55]. Such dependence of width on an essentially-flat landscape combined with the lack of explicit barrier crossing during the SGD descent forces us to rethink the analysis of basin hopping under a noisy dynamics.

Proposed framework: In this study, we aim at addressing these contradictions and come up with an arguably better-suited hypothesis for the stochastic gradient noise that has more pertinent theoretical implications for the phenomena associated with SGD. In particular, we go back to (3) and (4) and reconsider the application of CLT. This *classical* CLT assumes that U_k is a sum of many independent and identically distributed (i.i.d.) random variables, whose variance is *finite*, and then it states that the law of U_k converges to a Gaussian distribution, which then paves the way for (5). Even though the finite-variance assumption seems natural and intuitive at the first sight, it turns out that in many domains, such as turbulent motions [52], oceanic fluid flows [53], finance [35], biological evolution [26], audio signals [34], the assumption might fail to hold (see [13] for more examples). In such cases, the classical CLT along with the Gaussian approximation will no longer hold. While this might seem daunting, fortunately, one can prove an *extended* CLT and show that the law of the sum of these i.i.d. variables with infinite variance still converges to a family of *heavy-tailed* distributions that is called the α-stable distribution [32]. As we will detail in Section 2, these distributions are parametrized by their *tail-index* $\alpha \in (0, 2]$ and they coincide with the Gaussian distribution when $\alpha = 2$.

In this study, we relax the finite-variance assumption on the stochastic gradient noise and by invoking the extended CLT, we assume that U_k follows an α-stable distribution, as hinted in Figure 1(c). By following a similar rationale to (5) and (6), we reformulate SGD with this new assumption and consider its continuous-time limit for small step-sizes. Since the noise might not be Gaussian anymore (i.e. when $\alpha \neq 2$), the use of the Brownian motion would not be appropriate

in this case and we need to replace it with the α-stable Lévy motion, whose increments have an α-stable distribution [58]. Due to the heavy-tailed nature of α-stable distribution, the Lévy motion might incur large discontinuous jumps and therefore exhibits a fundamentally different behavior than the Brownian motion, whose paths are on the contrary almost surely continuous. As we will describe in detail in Section 2, the discontinuities also reflect in the metastability properties of Lévy-driven SDEs, which indicate that, as soon as $\alpha < 2$, the first exit time from a basin does *not* depend on its height; on the contrary, it directly depends on its width and the tail-index α. Informally, this implies that the process will *escape* from narrow minima – no matter how deep they are – and stay longer in wide minima. Besides, as α get smaller, the probability for the dynamics to jump in a wide basin will increase. Therefore, if the α-stable assumption on the stochastic gradient noise holds, then the existing metastability results automatically provide strong theoretical insights for illuminating the behavior of SGD.

Contributions: The main contributions of this paper are twofold: (i) we perform an extensive empirical analysis of the tail-index of the stochastic gradient noise in deep neural networks and (ii) based on these empirical results, we bring an alternative perspective to the existing approaches for analyzing SGD and shed more light on the folklore that SGD prefers wide minima by establishing a bridge between SGD and the related theoretical results from statistical physics and stochastic analysis.

We conduct experiments on the most common deep learning architectures. In particular, we investigate the tail behavior under fully-connected and convolutional models using negative log likelihood and linear hinge loss functions on MNIST, CIFAR10, and CIFAR100 datasets. For each configuration, we scale the size of the network and batch size used in SGD and monitor the effect of each of these settings on the tail index α.

Our experiments reveal several remarkable results:

- In all our configurations, the stochastic gradient noise turns out to be highly non-Gaussian and possesses a heavy-tailed behavior.
- Increasing the size of the minibatch has a very little impact on the tail-index, and as opposed to the common belief that larger minibatches result in Gaussian gradient noise, the noise is still far from being Gaussian.
- There is a strong interaction between the network architecture, network size, dataset, and the tail-index, which ultimately determine the dynamics of SGD on the training surface. This observation supports the view that, the geometry of the problem and the dynamics induced by the algorithm cannot be separated from each other.
- In almost all configurations, we observe two distinct phases of SGD throughout iterations. During the first phase, the tail-index rapidly decreases and SGD possesses a clear jump when the tail-index is at its lowest value and causes a sudden jump in the accuracy. This behavior strengthens the view that SGD crosses barriers at the very initial phase.

Our methodology also opens up several interesting future directions and open questions, as we discuss in Section 5.

2 Stable distributions and SGD as a Lévy-Driven SDE

The CLT states that the sum of i.i.d. random variables with a finite second moment converges to a normal distribution if the number of summands grow. However, if the variables have heavy-tail, the second moment may not exist. For instance, if their density $p(x)$ has a power-law tail decreasing as $1/|x|^{\alpha+1}$ where $0 < \alpha < 2$; only α-th moment exist with $\alpha < 2$. In this case, generalized central limit theorem (GCLT) says that the sum of such variables will converge to a distribution called the *α-stable* distribution instead as the number of summands grows (see e.g. [15]. In this work, we focus on the centered *symmetric α-stable* ($\mathcal{S}\alpha\mathcal{S}$) distribution, which is a special case of α-stable distributions that are symmetric around the origin.

We can view the $\mathcal{S}\alpha\mathcal{S}$ distribution as a heavy-tailed generalization of a centered Gaussian distribution. The $\mathcal{S}\alpha\mathcal{S}$ distributions are defined through their characteristic function via $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma) \iff \mathbb{E}[\exp(i\omega X)] = \exp(-|\sigma\omega|^{\alpha})$. Even though their probability density function does not admit a closed-form formula in general except in special cases, their density decays with a power law tail like $1/|x|^{\alpha+1}$ where $\alpha \in (0, 2]$ is called the *tail-index* which determines the behavior of the distribution: as α gets smaller; the distribution has a heavier tail. In fact, the parameter α also determines the moments: $\mathbb{E}[|X|^r] < \infty$ if and only if $r < \alpha$; implying X has infinite variance

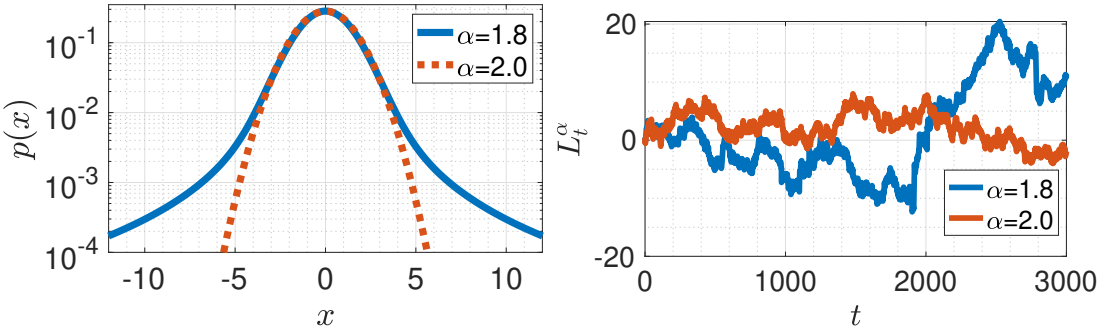


Figure 2: Left: $\mathcal{S}\alpha\mathcal{S}$ densities, right: L_t^α for $p = 1$. For $\alpha < 2$, $\mathcal{S}\alpha\mathcal{S}$ becomes heavier-tailed and L_t^α incurs jumps.

when $\alpha \neq 2$. The parameter $\sigma \in \mathbb{R}_+$ is known as the *scale* parameter and controls the spread of X around 0. We recover the Gaussian distribution $\mathcal{N}(0, 2\sigma^2)$ as a special case of $\mathcal{S}\alpha\mathcal{S}$ when $\alpha = 2$.

In this study, we make the following assumption on the stochastic gradient noise:

$$[U_k(\mathbf{w})]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{w})), \quad \forall i = 1, \dots, n \tag{7}$$

where $[v]_i$ denotes the i'th component of a vector v. Informally, we assume that each coordinate of U_k is $\mathcal{S}\alpha\mathcal{S}$ distributed with the same α and the scale parameter σ depends on the state $\mathbf{w}$. Here, this dependency is not crucial since we are mainly interested in the tail-index α, which can be estimated *independently* from the scale parameter. Therefore, we will simply denote $\sigma(\mathbf{w})$ as σ for clarity.

By using the assumption (7), we can rewrite the SGD recursion as follows:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k) + \eta^{1/\alpha} \left(\eta^{\frac{\alpha-1}{\alpha}} \sigma \right) S_k, \tag{8}$$

where $S_k \in \mathbb{R}^p$ is a random vector such that $[S_k]_i \sim \mathcal{S}\alpha\mathcal{S}(1)$. If the step-size η is small enough, then we can consider the continuous-time limit of this discrete-time process, which is expressed in the following SDE driven by an α-stable Lévy process:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \eta^{(\alpha-1)/\alpha} \sigma \, \mathrm{dL}_t^\alpha, \tag{9}$$

where L_t^α denotes the p-dimensional α-stable Lévy motion with *independent components*. In other words, each component of L_t^α is an independent α-stable Lévy motion in $\mathbb{R}$. For the scalar case it is defined as follows for $\alpha \in (0, 2]$ [13]:

(i) $\mathrm{L}_0^\alpha = 0$ almost surely.

(ii) For $t_0 < t_1 < \cdots < t_N$, the increments $(\mathrm{L}_{t_i}^\alpha - \mathrm{L}_{t_{i-1}}^\alpha)$ are independent $(i = 1, \dots, N)$.

(iii) The difference $(\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha)$ and L_{t-s}^α have the same distribution: $\mathcal{S}\alpha\mathcal{S}((t-s)^{1/\alpha})$ for $s < t$.

(iv) L_t^α is continuous in probability (i.e. it has *stochastically continuous* sample paths): for all $\delta > 0$ and $s \geq 0$, $p(|\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha| > \delta) \to 0$ as $t \to s$.

When $\alpha = 2$, L_t^α coincides with a scaled version of Brownian motion, $\sqrt{2}\mathrm{B}_t$. $\mathcal{S}\alpha\mathcal{S}$ and L_t^α are illustrated in Figure 2.

The SDE in (9) exhibits a fundamentally different behavior than the one in (6) does. This is mostly due to the stochastic continuity property of L_t^α, which enables L_t^α to have a countable number of discontinuities, which are sometimes called 'jumps'. In the rest of this section, we will recall important theoretical results about this SDE and discuss their implications on SGD.

For clarity of the presentation and notational simplicity we focus on the scalar case and consider the SDE (9) in $\mathbb{R}$ (i.e. $p = 1$). Multidimensional generalizations of the metastability results presented in this paper can be found in [23]. We rewrite (9) as follows:

$$\mathrm{d}w_t^\varepsilon = -\nabla f(w_t^\varepsilon)\mathrm{d}t + \varepsilon \mathrm{dL}_t^\alpha \tag{10}$$

for $t \geq 0$, started from the initial point $w_0 \in \mathbb{R}$, where L_t^α is the α-stable Lévy process, $\varepsilon \geq 0$ is a parameter and f is a non-convex objective with $r \geq 2$ local minima.

When $\varepsilon = 0$, we recover the gradient descent dynamics in continuous time: $\mathrm{d}w_t^0 = -\nabla f(w_t^0)\mathrm{d}t$, where the local minima are the stable points of this differential equation. However, as soon as

$\varepsilon > 0$, these states become 'metastable', meaning that there is a positive probability for w_t^ε to transition from one basin to another. However, the time required for transitioning to another basin strongly depends on the characteristics of the injected noise. The two most important cases are $\alpha = 2$ and $\alpha < 2$. When $\alpha = 2$, (i.e. the Gaussianity assumption) the process $(w_t^\varepsilon)_{t\geq 0}$ is continuous, which requires it to 'climb' the basin all the way up, in order to be able to transition to another basin. This fact makes the transition-time depend on the height of the basin. On the contrary, when $\alpha < 2$, the process can incur discontinuities and do not need to cross the boundaries of the basin in order to transition to another one since it can directly jump. This property is called the 'transition phenomenon' [13] and makes the transition-time mostly depend on the *width* of the basin. In the rest of the section, we will formalize these explanations.

Under some assumptions on the objective f, it is known that the process (10) admits a stationary density [48]. For a general f, an explicit formula for the equilibrium distribution is not known, however when the noise level ε is small enough, finer characterizations of the structure of the equilibrium density in dimension one is known. We next summarize known results in this area, which show that Lévy-driven dynamics spends more time in 'wide valleys' in the sense of [8] when ε goes to zero.

Assume that f is smooth with r local minima $\{m_i\}_{i=1}^r$ separated by $r-1$ local maxima $\{s_i\}_{i=1}^{r-1}$, i.e.

$$-\infty := s_0 < m_1 < s_1 < \cdots < s_{r-1} < m_r < s_r := \infty.$$

Furthermore, assume that the local minima and maxima are not degenerate, i.e. $f''(m_i) > 0$ and $f''(s_i) < 0$ for every i. We also assume the objective gradient has a growth condition $f'(w) > |w|^{1+c}$ for some constant $c > 0$ and when $|w|$ is large enough. Each local minima m_i lies in the (interval) valley $S_i = (s_{i-1}, s_i)$ of (width) length $L_i = |s_i - s_{i-1}|$. Consider also a δ-neighborhood $B_i := \{|x - m_i| \leq \delta\}$ around the local minimum with $\delta > 0$ small enough so that the neighborhood is contained in the valley S_i for every i. We are interested in the first exit time from B_i starting from a point $w_0 \in B_i$ and the transition time $T_{w_0}^i(\varepsilon) := \inf\{t \geq 0 : w_t^\varepsilon \notin \cup_{j\neq i} B_j\}$ to a neighborhood of another local minimum, we will remove the dependency to w_0 of the transition time in our discussions as it is clear from the context. The following result shows that the transition times are asymptotically exponentially distributed in the limit of small noise and scales like $\frac{1}{\varepsilon^\alpha}$ with ε.

Theorem 1 ([42]). *For an initial point $w_0 \in B_i$, in the limit $\varepsilon \to 0$, the following statements hold regarding the transition time:*

$$\begin{aligned} \mathbb{P}_{w_0}(T^i(\varepsilon) \in B_j) &\to q_{ij} q_i^{-1} \quad \text{if} \quad i \neq j, \\ \mathbb{P}_{w_0}(\varepsilon^\alpha T^i(\varepsilon) \geq u) &\leq e^{-q_i u} \quad \text{for any} \quad u \geq 0. \end{aligned}$$

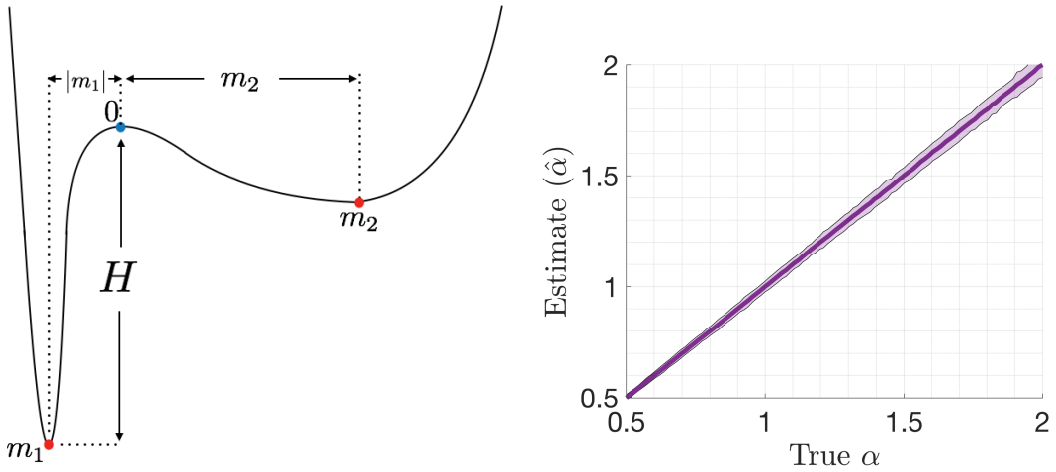
where

$$q_{ij} = \frac{1}{\alpha} \left| \frac{1}{|s_{j-1} - m_i|^\alpha} - \frac{1}{|s_j - m_i|^\alpha} \right|, \tag{11}$$

$$q_i = \sum_{j\neq i} q_{ij}. \tag{12}$$

If the SDE (10) would be driven by the Brownian motion instead, then an analogous theorem to Theorem 2 holds saying that the transition times are still exponentially distributed but the scaling ε^α needs to be replaced by e^{2H/ε^2} where H is the maximal depth of the basins to be traversed between the two local minima [7, 10]. This means that in the small noise limit, Brownian-motion driven gradient descent dynamics need exponential time to transit to another minimum whereas Levy-driven gradient descent dynamics need only polynomial time. We also note from Theorem 2 that the mean transition time between valleys for Lévy SDE does not depend on the depth H of the valleys they reside in which is an advantage over Brownian motion driven SDE in the existence of deep valleys. Informally, this difference is due to the fact that Brownian motion driven SDE has to typically climb up a valley to exit it, whereas Lévy-driven SDE could jump out.

The following theorem says that as $\varepsilon \to 0$, up to a normalization in time, the process w_t^ε behaves like a finite state-space Markov process that has support over the set of local minima $\{m_i\}_{i=1}^r$ admitting a stationary density $\pi = (\pi_i)_{i=1}^r$ with an infinitesimal generator Q. The process jumps between the valleys S_i, spending time proportional to probability p_i amount of time in each valley in the equilibrium where the probabilities $\pi = (\pi_i)_{i=1}^r$ are given by the solution to the linear system $Q\pi = 0$.



(a) (b)

Figure 3: (a) An objective with two local minima m_1, m_2 seperated by a local maxima at $s_1 = 0$. (b) Illustration of the tail-index estimator $\hat{\alpha}$.

Theorem 2 ([42]). *Let $w_0 \in S_i$, for some $1 \leq i \leq r$. For $t \geq 0$, $w^{\varepsilon}_{t\varepsilon^{-\alpha}} \to Y_{m_i}(t)$, as $\varepsilon \to 0$, in the sense of finite-dimensional distributions, where $Y = (Y_y(t))_{t\geq 0}$ is a continuous-time Markov chain on a state space $\{m_1, m_2, \ldots, m_r\}$ with the infinitesimal generator $Q = (q_{ij})_{i,j=1}^{r}$ with*

$$q_{ij} = \frac{1}{\alpha}\left|\frac{1}{|s_{j-1} - m_i|^{\alpha}} - \frac{1}{|s_j - m_i|^{\alpha}}\right|, \tag{13}$$

$$q_{ii} = -\sum\nolimits_{j\neq i} q_{ij}. \tag{14}$$

This process admits a density π satisfying $Q^T\pi = 0$.

A consequence of this theorem is that equilibrium probabilities p_i are typically larger for "wide valleys". To see this consider the special case illustrated in Figure 3(a) with $r = 2$ local minima $m_1 < s_1 = 0 < m_2$ separated by a local maximum at $s_1 = 0$. For this example, $m_2 > |m_1|$, and the second local minimum lies in a wider valley. A simple computation reveals

$$\pi_1 = \frac{|m_1|^{\alpha}}{|m_1|^{\alpha} + m_2^{\alpha}}, \quad \pi_2 = \frac{|m_2|^{\alpha}}{|m_1|^{\alpha} + |m_2|^{\alpha}}$$

We see that $\pi_2 > \pi_1$, that is in the equilibrium the process spends more time on the wider walley. In particular, the ratio $\frac{\pi_2}{\pi_1} = \left(\frac{m_2}{|m_1|}\right)^{\alpha}$ grows with an exponent α when the ratio $\frac{m_2}{|m_1|}$ of the width of the valleys grows. Consequently, if the gradient noise is indeed α-stable distributed, these results directly provide theoretical evidence for the wide-minima behavior of SGD.

3 Experimental Setup and Methodology

Experimental setup: We investigate the tail behavior of the stochastic gradient noise in a variety of scenarios. We first consider a fully-connected network (FCN) on the MNIST and CIFAR10 datasets. For this model, we vary the depth (i.e. the number of layers) in the set $\{2, 3, \ldots, 10\}$, the width (i.e. the number of neurons per layer) in the set $\{2, 4, 8, \ldots, 1024\}$, and the minibatch size ranging from 1 to full batch. We then consider a convolutional neural network (CNN) architecture (AlexNet) on the CIFAR10 and CIFAR100 datasets. We scale the number of filters in each convolutional layer in range $\{2, 4, \ldots, 512\}$. We randomly split the MNIST dataset into train and test parts of sizes 60K and 10K, and CIFAR10 and CIFAR100 datasets into train and test parts of sizes 50K and 10K, respectively. The order of the total number of parameters p range from several thousands to tens of millions.

For both fully connected and convolutional settings, we run each configuration with the negative-log-likelihood (i.e. cross entropy) and with the linear hinge loss, and we repeat each experiment with three different random seeds. The training algorithm is SGD with no explicit modification

such as momentum or weight decay. The training runs until 100% training accuracy is achieved or until maximum number of iterations limit is reached (the latter limit is effective in the under-parametrized models). At every 100th iteration, we log the full training and test accuracies, and the tail estimate of the gradients that are sampled using the corresponding mini-batch size. The codebase is implemented in python using pytorch and provided it in the supplementary material. Total runtime is ∼3 weeks on 8 relatively modern GPUs.

Method for tail-index estimation: Estimating the tail-index of an extreme-value distribution is a long-standing topic. Some of the well-known estimators for this task are [11, 12, 18, 43]. Despite their popularity, these methods are not specifically developed for α-stable distributions and it has been shown that they might fail for estimating the tail-index for α-stable distributions [38,41].

In this study, we use a relatively recent estimator proposed in [39] for α-stable distributions. It is given in the following theorem.

Theorem 3 ([39]). *Let $\{X_i\}_{i=1}^{K}$ be a collection of random variables with $X_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ and $K = K_1 \times K_2$. Define $Y_i \triangleq \sum_{j=1}^{K_1} X_{j+(i-1)K_1}$ for $i \in [\![1, K_2]\!]$. Then, the estimator*

$$\widehat{\frac{1}{\alpha}} \triangleq \frac{1}{\log K_1}\Big(\frac{1}{K_2}\sum_{i=1}^{K_2}\log|Y_i| - \frac{1}{K}\sum_{i=1}^{K}\log|X_i|\Big). \tag{15}$$

converges to $1/\alpha$ almost surely, as $K_2 \to \infty$.

As shown in Theorem 2.3 of [39], this estimator admits a provably faster convergence rate and smaller asymptotic variance than all the aforementioned methods.

In order to verify the accuracy of this estimator, we conduct a preliminary experiment, where we first generate $K = K_1 \times K_2$ many $\mathcal{S}\alpha\mathcal{S}(1)$ distributed random variables with $K_1 = 100$, $K_2 = 1000$ for 100 different values of α. Then, we estimate α by using $\hat{\alpha} \triangleq (\widehat{\frac{1}{\alpha}})^{-1}$. We repeat this experiment 100 times for each α. As shown in Figure 3(b), the estimator is very accurate for a large range of α. Due to its favorable theoretical properties such as independence of the scale parameter σ, combined with its empirical stability, we choose this estimator in our experiments.

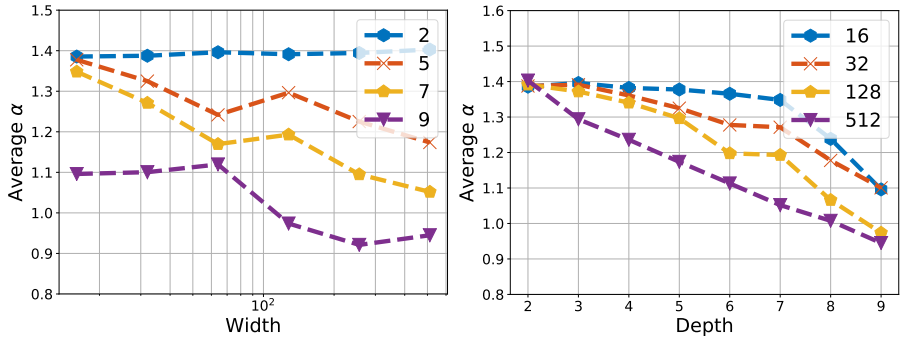
In order to estimate the tail-index α at iteration k, we first partition the set of data points $\mathcal{D} \triangleq \{1, \dots, n\}$ into many disjoint sets $\Omega_k^i \subset \mathcal{D}$ of size b, such that the union of these subsets give all the data points. Formally, for all $i, j = 1, \dots, n/b$, $|\Omega_k^i| = b$, $\cup_i \Omega_k^i = \mathcal{D}$, and $\Omega_k^i \cap \Omega_k^j = \emptyset$ for $i \neq j$. This approach is similar to sampling without replacement. We then compute the full gradient $\nabla f(\mathbf{w}_k)$ and the stochastic gradients $\nabla \tilde{f}_{\Omega_k^i}(\mathbf{w}_k)$ for each minibatch Ω_k^i. We finally compute the stochastic gradient noises $U_k^i(\mathbf{w}_k) = \nabla \tilde{f}_{\Omega_k^i}(\mathbf{w}_k) - \nabla f(\mathbf{w}_k)$, vectorize each $U_k^i(\mathbf{w}_k)$ and concatenate them to obtain a single vector, and compute the reciprocal of the estimator (15). In this case, we have $K = pn/b$ and we set K_1 to the divisor of K that is the closest to $\sqrt{K}$.

4 Results

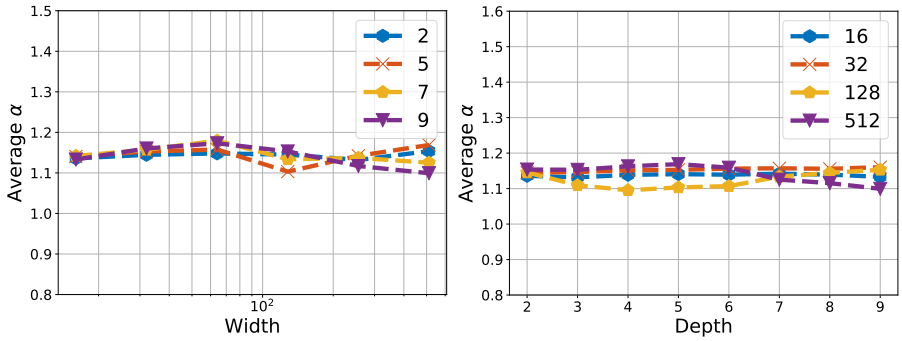
In this section we present the most important and representative results. We have observed that, in all configurations, the choice of the two loss functions and the three different initializations yield no significant difference. Therefore, throughout this section, we will focus on the negative-log-likelihood loss. Unless stated otherwise, we set the minibatch size $b = 500$ and the step-size $\eta = 0.1$.

Effect of varying network size: In our first set of experiments, we measure the tail-index for varying the widths and depths for the FCN, and varying widths (i.e. the number of filters) for the CNN. For very small sizes, the networks perform poorly, therefore, we only illustrate sufficiently large network sizes, which yield similar accuracies. For these experiments, we compute the average of the tail-index measurements for the last 10K iterations (i.e. when $\hat{\alpha}$ becomes stationary) to focus on the late stage dynamics.

Figure 4 shows the results for the FCN. The first striking observation is that in all the cases, the estimated tail-index is far from 2 with a very high confidence (the variance of the estimates were around 0.001), meaning that the distribution of the gradient noise is highly non-Gaussian. For the MNIST dataset, we observe that α systematically decreases for increasing network size, where this behavior becomes more prominent with the depth. This result shows that, for MNIST, increasing the dimension of the network results in a gradient noise with heavier tails and therefore increases the probability to end up in a wider basin.

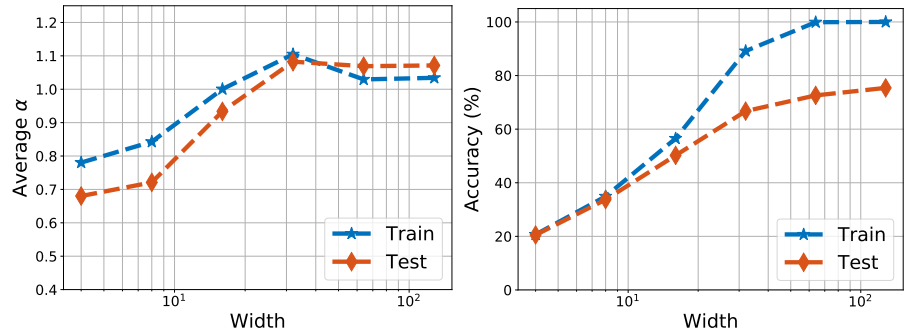


(a) MNIST

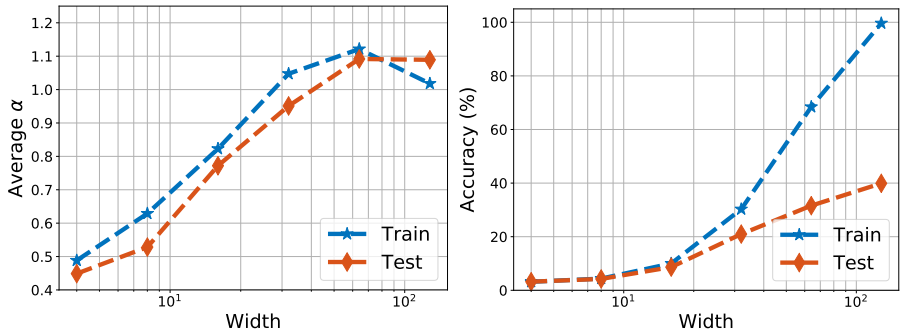


(b) CIFAR10

Figure 4: Estimation of α for varying widths and depths in FCN. The curves in the left figures correspond to different depths, and the ones on the right figures correspond to widths.



(a) CIFAR10

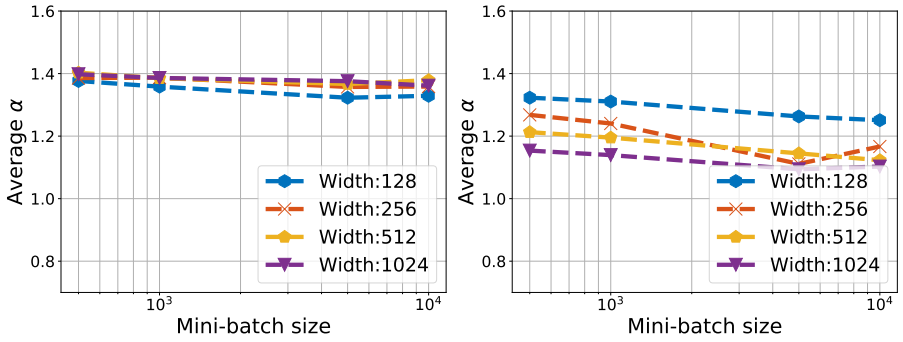


(b) CIFAR100

Figure 5: The accuracy and $\hat{\alpha}$ of the CNN for varying widths.

For the CIFAR10 dataset, we still observe that α is far from 2; however, in this case, increasing the network size does not have a clear effect on α: in all cases, we observe that α is in the range 1.1–1.2.

Figure 5 shows the results for the CNN. In this figure, we also depict the train and test accuracy, as well as the tail-index that is estimated on the test set. These results show that, for both CIFAR10



(a) Depth = 2 (b) Depth = 4

Figure 6: Estimation of α for varying minibatch size.

and CIFAR100, the tail-index is extremely low for the under-parametrized regime (e.g. the case when the width is 2, 4, or 8 for CIFAR10). As we increase the size of the network the value of α increases until the network performs reasonably well and stabilizes in the range 1.0–1.1. We also observe that α behaves similarly for both train and test sets[3].

These results show that there is strong interplay between the network architecture, dataset, and the algorithm dynamics: (i) we see that the size of the network can strongly influence α, (ii) for the exact same network architecture, the choice of the dataset has a significant impact on not only the landscape of the problem, but also the noise characteristics, hence on the algorithm dynamics.

Effect of the minibatch size: In our second set of experiments, we investigate the effect of the size of the minibatch on α. We focus on the FCN and monitor the behavior of α for different network and minibatch sizes b. Figure 6 illustrates the results. These rather remarkable results show that, as opposed to the common belief that the gradient noise behaves similar to a Gaussian for large b, the tail-index does not increase at all with the increasing b. We observe that α stays almost the same when the depth is 2 and it moves in a small interval when the depth is set to 4. We note that we obtained the same the train and test accuracies for different minibatch sizes.

Tail behavior throughout iterations: So far, we have focused on the last iterations of SGD, where α is in a stationary regime. In our last set of experiments, we shift our focus on the first iterations and report an interesting behavior that we observed in almost all our experiments. As a representative, in Figure 7, we show the temporal evolution of SGD for the FCN with 9 layers and 512 neurons/layer.

The results clearly show that there are two distinct phases of SGD (in this configuration before and after iteration 1000). In the first phase, the loss decreases very slowly, the accuracy slightly increases, and more interestingly α rapidly decreases. When α reaches its lowest level, the process possesses a jump, which causes a sudden decrease in the accuracy. After this point the process recovers again and we see a stationary behavior in α and an increasing behavior in the accuracy.

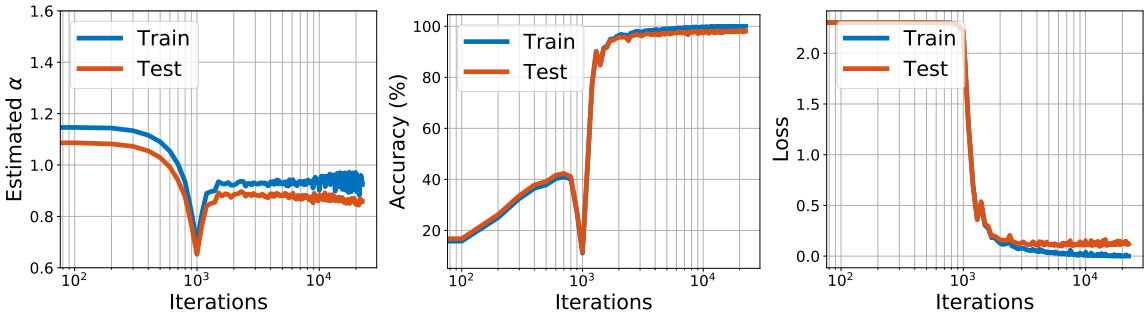
The fact that the process has a jump when α is at its smallest value provides a strong support to our assumptions and the metastability theory that we discussed in the previous section. Furthermore, these results further strengthen the view that SGD crosses barriers at the very initial phase. On the other hand, our current analysis is not able to determine whether the process jumps in a different basin or a 'better' part of the same basin and we leave it as a future work.

5 Conclusion and Open Problems

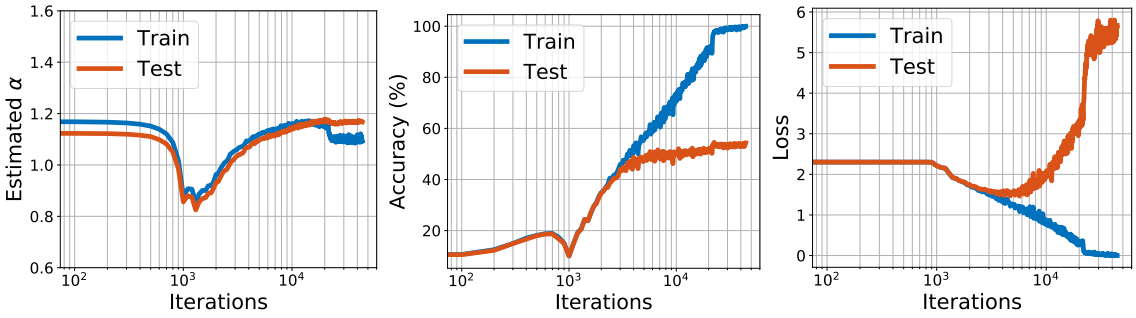
We investigated the tail behavior of the gradient noise in deep neural networks and empirically showed that the gradient noise is highly non-Gaussian. This outcome enabled us to analyze SGD as an SDE driven by a Lévy motion and establish a bridge between SGD and existing theoretical results, which provides more illumination on the behavior of SGD, especially in terms of choosing wide minima.

This study also brings up interesting open questions and future directions: (i) While the current metastability theory applies for the continuous-time processes, the behavior of the discretized process and its dependence on the algorithm parameters (e.g., the step-size, minibatch size) are

[3]We observed a similar behavior in under-parametrized FCN; however, did not plot those results to avoid clutter.



(a) MNIST



(b) CIFAR10

Figure 7: The iteration-wise behavior of of α for the FCN.

not clear and yet to be investigated. (ii) We observe that, especially during the first iterations, the tail-index depends on the current state $\mathbf{w}_k$, which suggests analyzing SGD as a stable-like process [3] where the tail-index can depend on time. However, the metastability behavior of these processes are not clear at the moment and its theory is still in an early phase [29]. (iii) Furthermore, an extension of the current metastability theory that includes minima with zero modes is also missing and appears to be challenging yet crucial direction of future research.

Acknowledgments

This work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014) project. Mert Gürbüzbalaban acknowledges support from the grants NSF DMS-1723085 and NSF CCF-1814888.

References

[1] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

[2] Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gerard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, volume 80, pages 314–323, Stockholm Sweden, 10–15 Jul 2018.

[3] R. F. Bass. Uniqueness in law for pure jump Markov processes. *Probability Theory and Related Fields*, 79(2):271–287, 1988.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.

[5] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[6] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

[7] Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

[8] P. Chaudhari, Anna Choromanska, S. Soatto, Yann LeCun, C. Baldassi, C. Borgs, J. Chayes, Levent Sagun, and R. Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, 2017.

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

[10] Martin V. Day. On the exponential exit law in the small parameter exit problem. *Stochastics*, 8(4):297–323, 1983.

[11] L. De Haan and L. Peng. Comparison of tail index estimators. *Statistica Neerlandica*, 52(1):60–70, 1998.

[12] A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

[13] J. Duan. *An Introduction to Stochastic Dynamics*. Cambridge University Press, New York, 2015.

[14] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *arXiv preprint arXiv:1507.05021*, 2015.

[15] Hans Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[16] M. I. Freidlin and A. D. Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

[17] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *arXiv preprint arXiv:1809.09349*, 2018.

[18] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, pages 1163–1174, 1975.

[19] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[21] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

[22] W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[23] P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in rd perturbed by multifractal Lévy noise. *Journal of Statistical Physics*, 141(1):94–119, 2010.

[24] P. Imkeller, I. Pavlyukevich, and T. Wetzel. The hierarchy of exit times of Lévy-driven Langevin equations. *The European Physical Journal Special Topics*, 191(1):211–222, Dec 2010.

[25] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[26] B. Jourdain, S. Méléard, and W. A. Woyczynski. Lévy flights in evolutionary ecology. *Journal of Mathematical Biology*, 65(4):677–707, 2012.

[27] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[29] I. Kuhwald and I. Pavlyukevich. Bistable behaviour of a jump-diffusion driven by a periodic stable-like additive process. *Discrete & Continuous Dynamical Systems-Series B*, 21(9), 2016.

[30] D. Lamberton and G. Pages. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. *Stochastics and dynamics*, 3(04):435–451, 2003.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436 EP –, 05 2015.

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

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

[34] A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2015.

[35] B. B. Mandelbrot. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E.* Springer Science & Business Media, 2013.

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

[37] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

[38] S. Mittnik and S. T. Rachev. Tail estimation of the stable index α. *Applied Mathematics Letters*, 9(3):53–56, 1996.

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

[40] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

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

[42] Ilya Pavlyukevich. Cooling down lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 40(41):12299, 2007.

[43] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[44] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703, 2017.

[45] G. O. Roberts and O. Stramer. Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, December 2002.

[46] Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *ICLR 2018 Workshop Contribution, arXiv:1706.04454*, 2017.

[47] Levent Sagun, V. Uğur Güney, Gérard Ben Arous, and Yann LeCun. Explorations on high dimensional landscapes. *International Conference on Learning Representations Workshop Contribution, arXiv:1412.6615*, 2015.

[48] G. Samorodnitsky and M. Grigoriu. Tails of solutions of certain nonlinear stochastic differential equations driven by heavy tailed Lévy motions. *Stochastic Processes and their Applications*, 105(1):69 – 97, 2003.

[49] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

[50] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[51] B. Tzen, T. Liang, and M. Raginsky. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In *Proceedings of the 2018 Conference on Learning Theory*, 2018.

[52] E. R. Weeks, T. H. Solomon, J. S. Urbach, and H. L. Swinney. Observation of anomalous diffusion and Lévy flights. In *Lévy flights and related topics in physics*, pages 51–71. Springer, 1995.

[53] W. A. Woyczyński. Lévy processes in the physical sciences. In *Lévy processes*, pages 241–266. Springer, 2001.

[54] Lei Wu, Chao Ma, and E Weinan. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8289–8298, 2018.

[55] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

[56] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3125–3136, 2018.

[57] S. Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

[58] V. V. Yanovsky, A. V. Chechkin, D. Schertzer, and A. V. Tur. Lévy anomalous diffusion and fractional Fokker–Planck equation. *Physica A: Statistical Mechanics and its Applications*, 282(1):13–34, 2000.

[59] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

[60] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1980–2022, 2017.

[61] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.