Multiplicative Noise and Heavy Tails in Stochastic Optimization

Liam Hodgkinson [1] **Michael W. Mahoney** [1]

Abstract

Although stochastic optimization is central to modern machine learning, the precise mechanisms underlying its success, and in particular, the precise role of the stochasticity, still remain unclear. Modeling stochastic optimization algorithms as discrete random recurrence relations, we show that multiplicative noise, as it commonly arises due to variance in local rates of convergence, results in heavy-tailed stationary behaviour in the parameters. Theoretical results are obtained characterizing this for a large class of (non-linear and even non-convex) models and optimizers (including momentum, Adam, and stochastic Newton), demonstrating that this phenomenon holds generally. We describe dependence on key factors, including step size, batch size, and data variability, all of which exhibit similar qualitative behavior to recent empirical results on state-of-the-art neural network models. Furthermore, we empirically illustrate how multiplicative noise and heavy-tailed structure improve capacity for basin hopping and exploration of non-convex loss surfaces, over commonly-considered stochastic dynamics with only additive noise and light-tailed structure.

1. Introduction

Relatively simple stochastic optimization procedures—in particular, those based on stochastic gradient descent (SGD)—have become the backbone of machine learning (ML) (Ma et al., 2018). To improve understanding of stochastic optimization in ML, and particularly why SGD and its extensions work so well, recent theoretical work has sought to study its properties and dynamics. Such analyses typically approach the problem through one of two perspectives. The first perspective, an *optimization (or quenching) perspective*, examines convergence either in expectation (Chen et al., 2019; Zhou et al., 2018; Gower et al., 2019; Nagaraj et al., 2019; Fontaine et al., 2020) or with some positive (high) probability (Roosta-Khorasani & Mahoney, 2016; Du et al., 2017; Kleinberg et al., 2018; Ward et al., 2019) through the lens of a deterministic counterpart. This perspective inherits some limitations of deterministic optimizers, including assumptions (e.g., convexity, Polyak-Łojasiewicz criterion, etc.) that are either not satisfied by state-of-the-art problems, or not strong enough to imply convergence to a quality (global) optimum. More concerning, however, is the inability to explain what has come to be known as the "generalization gap" phenomenon: increasing stochasticity by reducing batch size appears to improve generalization performance (Keskar et al., 2017; Martin & Mahoney, 2018). Empirically, existing strategies tend to break down for inference tasks when using large batch sizes (Golmant et al., 2018). The second perspective, a *probabilistic (annealing) perspective*, examines algorithms through the lens of Markov process theory (Freidlin & Wentzell, 1998; Henderson et al., 2003; Nemirovski et al., 2009). Here, stochastic optimizers are interpreted as samplers from probability distributions concentrated around optima, and annealing the optimizer (by reducing step size) increasingly concentrates probability mass around global optima. Traditional analyses trade restrictions on the objective for precise annealing schedules that guarantee adequate mixing and ensure convergence. However, it is uncommon in practice to consider step size schedules that decrease sufficiently slowly as to guarantee convergence to global optima with probability one (Li et al., 2020). In fact, SGD based methods with poor initialization can easily get stuck near poor local minima using a typical step decay schedule (Liu et al., 2019).

More recent efforts conduct a *distributional analysis*, directly examining the probability distribution that a stochastic optimizer targets for each fixed set of hyperparameters (Mandt et al., 2016; Babichev & Bach, 2018; Dieuleveut et al., 2017; Gürbüzbalaban et al., 2020). Here, one can assess a stochastic optimizer according to its capacity to reach and then occupy neighbourhoods of high-quality optima in the initial stages, where the step size is large and constant. As the step size is then rapidly reduced, tighter

[*]Equal contribution [1]ICSI and Department of Statistics, University of California, Berkeley, USA. Correspondence to: Liam Hodgkinson <liam.hodgkinson@berkeley.edu>.

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

neighbourhoods with higher probability mass surrounding nearby minima are achievable. This is most easily accomplished using a variational approach by appealing to continuous-time Langevin approximations (Mandt et al., 2016; Chaudhari & Soatto, 2018), whose stationary distributions are known explicitly (Ma et al., 2015). However, these approaches also require restrictive assumptions, such as constant or bounded volatility (Mandt et al., 2017; Orvieto & Lucchi, 2019). Interestingly, these assumptions parallel the common belief that the predominant part of the stochastic component of an optimizer is an additive perturbation (Kleinberg et al., 2018; Zhang et al., 2019a). Such analyses have been questioned with recent discoveries of non-Gaussian noise (Şimşekli et al., 2019; Gürbüzbalaban et al., 2020) that leads to *heavy-tailed* stationary behaviour (i.e., not *light-tailed*, where distributions have finite Laplace transform). This behavior implies stronger exploratory properties and an increased tendency to rapidly reach faraway basins than earlier Langevin-centric analyses suggest. In particular, Şimşekli et al. (2020a) used fractal dimension theory to show that if such heavy-tailed exploration is present, then it can improve generalization, with test accuracies typically increasing when tails are heavier.

Main Contributions. We model stochastic optimizers as Markov random recurrence relations, thereby avoiding continuous-time approximations, and we examine their stationary distributions. The formulation of this model is described in §2. We show that *multiplicative noise*, frequently assumed away in favour of more convenient *additive noise* in continuous analyses, can often lead to heavy-tailed stationary behaviour. This plays a critical role in the dynamics of a stochastic optimizer, and it influences the capacity of the algorithm to hop between basins in the loss landscape. In this paper, we consider heavy-tailed behavior that assumes a power law functional form. We say that the stationary distributions of the parameters/weights W exhibit *power laws*, with tail probabilities $\mathbb{P}(\|W\| > t) = \Omega(t^{-\alpha})$ as $t \to \infty$, for some $\alpha > 0$ called the *tail exponent* (where smaller tail exponents correspond to heavier tails). Further details, including precise definitions, are in Appendix A.

To inform our analysis, in §3, we consider the special case of constant step-size SGD applied to linear least squares, which obeys a *random linear recurrence relation* displaying both multiplicative and additive noise. Using well-known results (Buraczewski et al., 2016), we isolate three regimes determining the tail behaviour of SGD (shown in Table 1, discussed in §3), finding stationary behaviour *always* exhibits a precise power law in an infinite data regime.

In §4, we extend these observations by providing sufficient conditions for the existence of power laws arising in arbitrary iterative stochastic optimization algorithms, on both convex and non-convex problems, with more precise results when updates are bilipschitz. Factors influencing tail behaviour are examined, with existing empirical findings supporting the hypothesis that heavier tails coincide with improved generalization performance.

Numerical experiments are conducted in §5, illustrating how multiplicative noise and heavy-tailed stationary behaviour improve the capacity for basin hopping (relative to light-tailed stationary behaviour) in the exploratory phase of learning. We finish by discussing impact on related work in §6, including a continuous-time analogue of Table 1 (shown in Table 2).

Related Work. There is a large body of related work, and we review only the most directly related. Analysis of stochastic optimizers via stationary distributions of Markov processes was recently considered in Mandt et al. (2016); Babichev & Bach (2018); Dieuleveut et al. (2017). The latter, in particular, examined first and second moments of the stationary distribution, although these can be ineffective measures of concentration in heavy-tailed settings. Heavy tails in ML have been observed and empirically examined in spectral distributions of weights (Martin & Mahoney, 2017; 2018; 2019; 2020a;b) and in the weights themselves (Şimşekli et al., 2019; Panigrahi et al., 2019; Şimşekli et al., 2020a), but (ML style) theoretical analyses seem limited to continuous-time examinations (Şimşekli et al., 2019; 2020b). Connections between multiplicative noise and heavy-tailed fluctuations can be seen throughout the wider literature (Deutsch, 1993; Frisch & Sornette, 1997; Sornette & Cont, 1997; Buraczewski et al., 2016). From a physical point of view, multiplicative noise acts as an external environmental field, capable of exhibiting drift towards low energy states (Volpe & Wehr, 2016). Hysteretic optimization (Pál, 2006) is one example of a stochastic optimization algorithm taking advantage of this property. Closest to our own work is Gürbüzbalaban et al. (2020), which conducts a detailed analysis of heavy tails in the stationary distribution of SGD applied to least-squares linear regression. Our main objective in this paper is to provide far-reaching extensions of these results to more general stochastic optimization algorithms and problems. To our knowledge, no other theoretical analysis of this phenomenon has been conducted in a general optimization setting. Indeed, while multiplicative noise in stochastic optimization has been explored in some recent empirical analyses (Wu et al., 2020; Zhang et al., 2018; Holland, 2019), and its presence documented (Xing et al., 2018), its impact appears underappreciated, relative to the well-studied and exploited effects of additive noise (Ge et al., 2015; Jin et al., 2017; Du et al., 2017; Kleinberg et al., 2018). Section 6 contains a discussion of additional related work in light of our results.

Regime	Condition on A	Tails for B	Tails for W_∞
Light-tailed (§A.1)	$\mathbb{P}(\|A\| \leqslant 1) = 1$	$\mathcal{O}(e^{-\lambda t})$	$\mathcal{O}(e^{-\rho t})$
Heavy-tailed (multiplicative) (§A.2)	$\|\sigma_{\min}(A)\|_\alpha = 1$	$o(t^{-\alpha})$	$\Omega(t^{-\alpha})$
Heavy-tailed (additive) (§A.3)	—	$\Omega(t^{-\beta})$	$\Omega(t^{-\beta})$

Table 1: Summary of the three primary tail behaviour regimes for the stationary distribution of (5).

Regime	Noise	Condition on r	Quadratic Example
Light-tailed (Ma et al., 2015)	White noise	$\liminf_{\|w\|\to\infty} r(w) > 0$	g bounded
Heavy-tailed (multiplicative) (Ma et al., 2015)	White noise	$\limsup_{\|w\|\to\infty} r(w) = 0$	$g(w) = c\nabla f(w)$
Heavy-tailed (additive) (Şimşekli et al., 2019)	Lèvy noise	—	arbitrary g

Table 2: Summary of three tail behaviour regimes for continuous-time models (8).

Notation. We let I_d denote the $d \times d$ identity matrix. Unless otherwise stated, we default to the following norms: for vector x, we let $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$ denote the Euclidean norm of x, and for matrix A, the norm $\|A\| = \sup_{\|x\|=1} \|Ax\|$ denotes the ℓ^2-induced (spectral) norm. Also, for matrix A, $\|A\|_F = \|\mathrm{vec}(A)\|$ is the Frobenius norm, and $\sigma_{\min}(A)$, $\sigma_{\max}(A)$ are its smallest and largest singular values, respectively. For two vectors/matrices A, B, we let $A \otimes B$ denote their tensor (or Kronecker) product. For $x \in \mathbb{R}$, $\log^+ x = \log\max\{1, x\}$. Knuth asymptotic notation (Knuth, 1976) is adopted. The notation $(\Omega, \mathcal{E}, \mathbb{P})$ is used to denote an appropriate underlying probability space. For two random elements X, Y, $X \overset{\mathcal{D}}{=} Y$ if X, Y have the same distribution. Finally, for random vector/matrix X, for $\alpha > 0$, $\|X\|_\alpha = (\mathbb{E}\|X\|^\alpha)^{1/\alpha}$. All proofs are relegated to Appendix D.

2. Stochastic optimization as a Markov chain

In this section, we describe how to model a stochastic optimization algorithm as a Markov chain — in particular, as a *random recurrence relation*. This formulation is uncommon in the ML literature, but will be important for our analysis. Consider a general single-objective stochastic optimization problem, where the goal is to solve problems of the form $w^* = \arg\min_w \mathbb{E}_{\mathcal{D}} \ell(w, X)$ for some scalar loss function ℓ, and random element $X \sim \mathcal{D}$ (the data) (Kroese et al., 2013, §12). In the sequel, we shall assume the weights w occupy a vector space S with norm $\|\cdot\|$. To minimize ℓ with respect to w, some form of fixed point iteration is typically adopted. Supposing that there exists some continuous map Ψ such that any fixed point of $\mathbb{E}\Psi(\cdot, X)$ is a minimizer of ℓ, the sequence of iterations

$$W_{k+1} = \mathbb{E}_{\mathcal{D}} \Psi(W_k, X) \tag{1}$$

either diverges, or converges to a minimizer of ℓ (Granas & Dugundji, 2013). In practice, this expectation may not be easily computed, and so one could instead consider the sequence of iterated random functions: for $X_i^{(k)} \overset{\text{iid}}{\sim} \mathcal{D}$,

$$W_{k+1} = \Psi(W_k, X_k) \qquad k = 0, 1, \ldots \tag{2}$$

For example, one could consider the Monte Carlo approximation of (1):

$$W_{k+1} = \frac{1}{n} \sum_{i=1}^n \Psi(W_k, X_i^{(k)}), \tag{3}$$

which can be interpreted in the context of *randomized subsampling with replacement*, where each $(X_i^{(k)})_{i=1}^n$ is a batch of n subsamples drawn from a large dataset $\mathcal{D}$. Subsampling without replacement can also be treated in the form (3) via the Markov chain $\{W_{sk}\}_{k=0}^\infty$, where s is the number of minibatches in each epoch.

Since (2) will not, in general, converge to a single point, the *stochastic approximation (SA)* approach of Robbins & Monro (1951) considers a corresponding sequence of maps Ψ_k given by $\Psi_k(w, x) := (1 - \gamma_k)w + \gamma_k \Psi(w, x)$, where $\{\gamma_k\}_{k=1}^\infty$ is a decreasing sequence of *step sizes*. Provided Ψ is uniformly bounded, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$, the sequence of iterates $W_{k+1} = n^{-1} \sum_{i=1}^n \Psi_k(W_k, X_i^{(k)})$ converges in L^2 and almost surely to a minimizer of ℓ (Blum, 1954; Nemirovski et al., 2009). Note that this differs from the *sample average approximation (SAA)* approach (Kleywegt et al., 2002), where a deterministic optimization algorithm is applied to a random objective (e.g., gradient descent on subsampled empirical risk).

Here are examples of how popular ML stochastic optimization algorithms fit into this framework.

Example 1 (SGD & SGD with momentum)**.** Minibatch SGD with step size γ coincides with (3) under $\Psi(w, X) = w - \gamma\nabla\ell(w, X)$. Incorporating momentum is possible by considering the augmented space of weights and velocities together. In the standard setup, letting v denote velocity, and w the weights, $\Psi((v, w), X) = (\eta v + \nabla\ell(w, X), w - \gamma(\eta v + \nabla\ell(w, X)))$.

Example 2 (Adam)**.** Using state-space augmentation, the popular first-order Adam optimizer (Kingma & Ba, 2015) can also be cast into the form of (2). Indeed, for $\beta_1, \beta_2, \epsilon > 0$, letting $g(w, X) = \sum_{i=1}^n \nabla \ell(w, X_i)$, for $M(m, w, X) = \beta_1 m + n^{-1}(1-\beta_1) g(w, X)$, $V(v, w, X) = \beta_2 v + (1-\beta_2) n^{-2} g(w, X)^2$, $B_1(b) = 1 - \beta_1(1-b)$, $B_2(b) = 1 - \beta_2(1-b)$, and

$$W(b_1, b_2, m, v, w, X) = w - \eta \frac{M(m, w, X)/B_1(b_1)}{\sqrt{V(v, w, X)/B_2(b_2)} + \epsilon},$$

iterations of the Adam optimizer satisfy $\mathcal{W}_{k+1} = \Psi(\mathcal{W}_k, X_k)$, where $\Psi((b_1, b_2, m, v, w), X) = (B_1(b_1), B_2(b_2), M(m, w, X), V(v, w, X), W(b_1, b_2, m, v, w, X))$.

Example 3 (Stochastic Newton)**.** The formulation (2) is not limited to first-order stochastic optimization. Indeed, for $g(w, X) = \sum_{i=1}^n \nabla \ell(w, X_i)$ and $H(w, X) = \sum_{i=1}^n \nabla^2 \ell(w, X_i)$, the choice $\Psi(w, X) = w - \gamma H(w, X)^{-1} g(w, X)$ coincides with the stochastic Newton method (Roosta-Khorasani & Mahoney, 2016).

The iterations (2) are also sufficiently general to incorporate other random effects, e.g., dropout (Srivastava et al., 2014). Instead of taking the SA approach, in our analysis we will examine the original Markov chain (2), where hyperparameters (including step size) are fixed. This will provide a clearer picture of ongoing dynamics at each stage within an arbitrary learning rate schedule (Babichev & Bach, 2018). Also, assuming reasonably rapid mixing, global tendencies of a stochastic optimization algorithm, such as the degree of exploration in space, can be examined through the *stationary distribution* of (2).

3. The linear case with SGD

As a warmup to our main results, let us first consider the special case of ridge regression, i.e., least squares linear regression with L^2 regularization, using vanilla SGD, as treated in Gürbüzbalaban et al. (2020). Let $\mathcal{D}$ denote a dataset comprised of inputs $x_i \in \mathbb{R}^d$ and corresponding labels $y_i \in \mathbb{R}^m$. Supposing that (X, Y) is a pair of random vectors uniformly drawn from $\mathcal{D}$, for $\lambda \geqslant 0$, we seek a solution to

$$M^* = \operatorname*{arg\,min}_{M \in \mathbb{R}^{d \times m}} \tfrac{1}{2} \mathbb{E}_{\mathcal{D}} \|Y - MX\|^2 + \tfrac{1}{2} \lambda \|M\|_F^2. \qquad (4)$$

Applying minibatch SGD to solving (4) with constant (sufficiently small) step size results in a Markov chain $\{M_k\}_{k=0}^\infty$ in the estimated parameter matrix. We start with the following immediate, but important, observation.

Lemma 1. *Let n denote the size of each minibatch $(X_{ik}, Y_{ik})_{i=1}^n$ comprised of independent and identically distributed copies of (X, Y) for $k = 1, 2, \ldots$. For W_k the vectorization of M_k, iterations of minibatch SGD undergo the following random linear recurrence relation*

$$W_{k+1} = A_k W_k + B_k, \qquad (5)$$

where $A_k = I_m \otimes ((1-\lambda\gamma) I_d - \gamma n^{-1} \sum_{i=1}^n X_{ik} X_{ik}^\top)$, and $B_k = \gamma n^{-1} \sum_{i=1}^n Y_{ik} \otimes X_{ik}$. If A_k, B_k are non-atomic, $\log^+ \|A_k\|$, $\log^+ \|B_k\|$ are integrable, and $\mathbb{E}_{\mathcal{D}} \log \|A_k\| < 0$, then (5) is ergodic.

Note that SGD possesses *multiplicative noise* in the form of the factor A_k, as well as nonzero *additive noise* in the form of the factor B_k. Under the conditions of Lemma 1, the expectations of (5) converge to M^*. Although the dynamics of this process, as well as the shape of its stationary distribution, are not as straightforward, random linear recurrence relations are among the most well-studied discrete-time processes, and as multiplicative processes, are well-known to exhibit heavy-tailed behaviour. In Appendix A, we discuss classical results on the topic. The three primary tail regimes are summarized in Table 1, where each $\alpha, \beta, \lambda, \rho$ denotes a strictly positive value. There are two possible mechanisms by which the stationary distribution of (5) can be heavy-tailed. Most discussions about SGD focus on the additive noise component B_k (Kleinberg et al., 2018). In this case, if B_k is heavy-tailed, then the stationary distribution of (5) is also heavy-tailed. This is the assumption considered in Şimşekli et al. (2019; 2020b). However, this is not the only way heavy-tailed noise can arise. In fact, we have the following result.

Lemma 2. *Assuming the distribution of X has full support on $\mathbb{R}^d$, there exists $\alpha > 0$ such that $\|\sigma_{\min}(A_k)\|_\alpha = 1$. If (5) is ergodic, its stationary distribution is heavy-tailed with tail exponent at most α, that is, $\mathbb{P}(\|W_\infty\| > t) \geqslant C_\alpha (1+t)^{-\alpha}$ for some $C_\alpha > 0$ and any $t \geqslant 0$.*

Lemma 2 suggests that heavy-tailed fluctuations could be more common than previously considered. This is perhaps surprising, given that common Langevin approximations of SGD applied to this problem exhibit light-tailed stationary distributions (Mandt et al., 2016; Orvieto & Lucchi, 2019). A basic reasoning behind Lemma 2 is as follows: for $\beta > \alpha$, $\|\sigma_{\min}(A_k)\|_\beta > 1$, so iterating (5) implies $\|W_k\|_\beta \to \infty$, hence the limiting distribution must be heavy-tailed. The multiplicative heavy-tailed regime illustrates how a power law can arise from light-tailed data (in fact, even from data with finite support). Important to note here is that if B_k displays (not too significant) heavy-tailed behaviour, the tail exponent of the stationary distribution is entirely due to the recursion and properties of the multiplicative noise A_k. Here, the heavy-tailed behaviour arises due to *intrinsic factors* to the stochastic optimization, and they tend to dominate over time. Similar observations were reported in Gürbüzbalaban et al. (2020), followed by a thorough investigation into the tail exponent α. Increas-

ing step size, decreasing batch size, and increasing dimension were all shown to result in heavier tails. Our primary objective is to extend these ideas to more general settings.

4. Power laws for general objectives and optimizers

In this section, we consider the general case (2), and we examine how heavy-tailed stationary behaviour can arise for any stochastic optimizer, applied to both convex and non-convex problems. As in Section 3, here we are most interested in the presence of heavy-tailed fluctuations due to multiplicative factors. The case when heavy-tailed fluctuations arise from the additive noise case is clear: if, for every $w \in \mathbb{R}^d$, $\Psi(w, X)$ is heavy-tailed/has infinite αth-moment, then for each $k = 1, 2, \ldots$, W_k is heavy-tailed/has infinite αth moment also, irrespective of the dynamics of W.

In our main result (Theorem 1), we illustrate how power laws arise in general smooth stochastic optimization algorithms that are contracting on average and also strongly convex near infinity with positive probability.

Theorem 1. *Let $(S, \|\cdot\|)$ be a separable Banach space. Assume that $\Psi : S \times \Omega \to S$ is a random function on S such that Ψ is a.s. Lipschitz and has probability measure with a positively supported absolutely continuous component with respect to a σ-finite non-null probability measure on S. Let K_Ψ denote a random variable such that $K_\Psi(\omega)$ is the Lipschitz constant of $\Psi(\cdot, \omega)$ for each $\omega \in \Omega$. Assume that K_Ψ is integrable, and $\|\Psi(w) - w\|$ is integrable for some $w^* \in S$. Suppose there exist non-negative random variables k_Ψ, M_Ψ such that, almost surely, for all $w \in S$,*

$$k_\Psi \|w - w^*\| - M_\Psi \leqslant \|\Psi(w) - \Psi(w^*)\| \leqslant K_\Psi \|w - w^*\|. \tag{6}$$

If $\mathbb{E} \log K_\Psi < 0$, then the Markov chain given by $W_{k+1} = \Psi_k(W_k), k = 0, 1, \ldots,$ where each Ψ_k is an independent and identically distributed copy of Ψ, is geometrically ergodic with $W_\infty = \lim_{k\to\infty} W_k$ satisfying the distributional fixed point equation $W_\infty \overset{\mathcal{D}}{=} \Psi(W_\infty)$. Furthermore, if $k_\Psi > 1$ with positive probability and $\|\Psi(w^*)\|_\alpha < \infty$ for any $\alpha > 0$, then:

1. *There exist $\mu, \nu, C_\mu, C_\nu > 0$ such that $C_\mu(1 + t)^{-\mu} \leqslant \mathbb{P}(\|W_\infty\| > t) \leqslant C_\nu t^{-\nu}$, for all $t > 0$.*
2. *There exist $\alpha, \beta > 0$ such that $\|K_\Psi\|_\beta = 1$ and $\|k_\Psi\|_\alpha = 1$ and for $\delta > 0$, $0 < \limsup_{t\to\infty} t^{\alpha+\epsilon}\mathbb{P}(\|W_\infty\| > t)$, and $\limsup_{t\to\infty} t^{\beta-\epsilon}\mathbb{P}(\|W_\infty\| > t) < \infty$.*

Geometric rates of convergence in the Prohorov and total variation metrics to stationarity are discussed in Diaconis & Freedman (1999); Alsmeyer (2003). From Theorem 1, we find that the presence of expanding multiplicative noise implies the stationary distribution of (2) is stochastically bounded between two power laws. This suggests that smooth stochastic optimizers satisfying (6) and $\mathbb{P}(k_\Psi > 1) > 0$ can conduct wide exploration of the loss landscape. To our knowledge, these conditions for heavy-tailed stationary behaviour are significantly weaker than the present literature suggests (Gürbüzbalaban et al., 2020; Mirek, 2011; Buraczewski et al., 2016).

Example 4 (Heavy tails in SGD)**.** For example, condition (6) holds for SGD for any loss ℓ that is strongly convex outside of a bounded region (e.g. when weight decay is added). In this case,

$$k_\Psi = \liminf_{\|w\|\to\infty} \sigma_{\min}(I - \gamma\nabla^2\ell(w, X)), \text{ and,} \tag{7a}$$

$$K_\Psi = \sup_w \|I - \gamma\nabla^2\ell(w, X)\|. \tag{7b}$$

In the linear case (5), k_Ψ reduces to $\sigma_{\min}(A)$, hence, using Theorem 1, we can recover Lemma 2. Extending Lemma 2 to the general case, vanilla SGD exhibits heavy-tailed behaviour when $\nabla^2\ell(w, X)$ has unbounded spectral distribution (as might be the case when X has full support on $\mathbb{R}^d$).

Conditions for stochastic Newton are similar, albeit more complex. In particular, stochastic Newton exhibits heavy-tailed behaviour when the Jacobian of $H(w, X)^{-1}g(w, X)$ has unbounded spectral distribution. Adaptive optimizers such as momentum and Adam incorporate geometric decay that can prevent heavy-tailed fluctuations, potentially limiting exploration while excelling at exploitation to nearby optima. It has been suggested that these adaptive aspects should be turned off during an initial warmup phase (Liu et al., 2020), implying that exacerbating heavy-tailed fluctuations and increased tail exponents could aid exploratory behaviour in the initial stages of training. On the other hand, if heavy-tailed behaviour is so extreme as to be detrimental for exploiting nearby optima, adaptive optimizers can prove effective (Zhang et al., 2019b).

To treat other stochastic optimizers that are not Lipschitz, or do not satisfy the conditions of Theorem 1, we present in Lemma 3 an abstract sufficient condition for heavy-tailed stationary distributions of ergodic Markov chains.

Lemma 3. *For a general metric space S, suppose that W is an ergodic Markov chain on S with $W_k \overset{\mathcal{D}}{\to} W_\infty$ as $k \to \infty$, and let $f : S \to \mathbb{R}$ be some scalar-valued function. If there exists some $\epsilon > 0$ such that $\inf_{w\in S} \mathbb{P}(|f(\Psi(w))| > (1 + \epsilon)|f(w)|) > 0$, then $f(W_\infty)$ is heavy-tailed.*

Under our model, Lemma 3 says that convergent constant step-size stochastic optimizers will exhibit heavy-tailed stationary behaviour if there is some positive probability of the optimizer moving further away from an optimum, irrespective of how near or far you are from it. To our knowledge, this is the first time that variational consequences of this property have been considered in stochas-

tic optimization. Analyses concerning stochastic optimization algorithms typically consider rates of contraction towards a nearby optimum, quantifying the *exploitative* behaviour of a stochastic optimizer. In the convex setting, that stochastic optimization algorithms could, at any stage, move away from any optimum, appears detrimental, but is critical in non-convex settings, where *exploration* (rather than exploitation to a local optimum) is important. Indeed, we find that it is this behaviour that directly determines the tails of the stochastic optimizer's stationary distribution, and therefore, its exploratory behaviour.

Factors influencing the tail exponent. Using bounds from Theorem 1, we may extend observations in Gürbüzbalaban et al. (2020) to the general case. According to Theorem 1, the following factors play a role in *decreasing the tail exponent* (generally, increasing k_Ψ, K_Ψ in expectation or dispersion implies decreased α) resulting in *heavier-tailed noise*. In each case (excluding step size, which is complicated since step size also affects stability), the literature supports the hypothesis that factors influencing heavier tails coincide with improved generalization performance; see Martin & Mahoney (2018); Jastrzębski et al. (2018) and references therein.

Decreasing batch size / increasing step size: As the mini-batch size $n \to \infty$ or step size $\gamma \to 0$, $\mathrm{Var}(k_\Psi) \to 0$ and so $\mathbb{P}(k_\Psi > 1) \to 0$ if W is ergodic. Conversely, by Theorem 1, decreasing batch size or increasing step size results in an decreased (i.e., heavier) tail exponent for ergodic stochastic optimizers, keeping in mind that step sizes also affect stability of the algorithm. This is in line with Jastrzębski et al. (2018); Yao et al. (2018) and it suggests a relationship to the *generalization gap* phenomenon. The relationship between step and batch sizes has received attention (Balles et al., 2017; Smith et al., 2018); choosing these parameters to increase heavy-tailed fluctuations while keeping variation sensible could yield a valuable exploration strategy.

More dispersed data: Increasing dispersion in the data implies increased dispersion for the distribution of k_Ψ, K_Ψ, and hence heavier tails. For the same model trained to different datasets, a smaller tail exponent may be indicative of richer data, not necessarily of higher variance, but exhibiting a larger moment of some order. Data augmentation is a strategy to achieve this (Wang & Perez, 2017).

Increasing regularization: Provided W remains ergodic, the addition of a large explicit L^2-regularizer to the objective function (known to help avoiding bad minima (Liu et al., 2019)) results in larger k_Ψ, and hence, heavier-tailed noise.

Increasing dimension: The effect of dimension is noteworthy for ML, albeit not so straightforward at higher generality, where direct comparison between model classes becomes complex. For neural networks in the SGD case, a Wishart+Wigner Hessian model (Pennington & Bahri, 2017) together with (7a) suggests that k_Ψ should increase (and the tail exponent decrease) with dimension — see Appendix B for details. Later in Figure 3, we show empirically that increasing depth within the same architecture class also yields heavier tails.

5. Numerical Experiments

To illustrate the advantages that multiplicative noise and heavy tails offer in non-convex optimization, in particular for exploration (of the entire loss surface) versus exploitation (to a local optimum), we first consider stochastic optimization algorithms in the non-stationary regime. To begin, in one dimension, we compare perturbed gradient descent (GD) with additive light-tailed and heavy-tailed noise against a version with additional multiplicative noise. That is,

$$\text{(a,b)} \quad w_{k+1} = w_k - \gamma(f'(w_k) + (1+\sigma)Z_k),$$
$$\text{(c)} \quad w_{k+1} = w_k - \gamma((1+\sigma Z_k^{(1)})f'(w_k) + Z_k^{(2)}),$$

where $f : \mathbb{R} \to \mathbb{R}$ is the objective function, $\gamma, \sigma > 0$, $Z_k, Z_k^{(1)}, Z_k^{(2)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ in (a) and (c), and in (b), Z_k are i.i.d. t-distributed (heavy-tailed) with 3 degrees of freedom, normalized to have unit variance. Algorithms (a)-(c) are applied to the objective $f(x) = \frac{1}{10}x^2 + 1 - \cos(x^2)$, which achieves a global (wide) minimum at zero. Iterations of (a)-(c) have expectation on par with classical GD. For fixed step size $\gamma = 10^{-2}$ and initial $w_0 = -4.75$, the distribution of 10^6 successive iterates are presented in Figure 1 for small ($\sigma = 2$), moderate ($\sigma = 12$), and strong ($\sigma = 50$) noise. Both (b) and (c) readily enable jumps between basins. However, while the additive noise optimizers (a), (b) smooth the effective loss landscape, making it easier to jump between basins, it also has the side effect of reducing resolution in the vicinity of minima. On the other hand, the multiplicative noise optimizer (c) maintains close proximity (peaks in the distribution) to critical points.

Figure 1 also illustrates exploration/exploitation benefits of multiplicative noise (and associated heavier tails) over additive noise (and associated lighter tails). While rapid exploration may be achieved using heavy-tailed additive noise (Şimşekli et al., 2019), since reducing the step size may not reduce the tail exponent, efficient exploitation of local minima can become challenging, even for small step sizes. On the other hand, multiplicative noise has the benefit of behaving similarly to Gaussian additive noise for small step sizes (Rubin et al., 2014). We can see evidence of this behaviour in the leftmost column in Figure 1, where the size of the multiplicative noise is small. As the step size is annealed to zero, multiplicative noise resembles the convolutional nature of additive noise (Kleinberg et al., 2018).

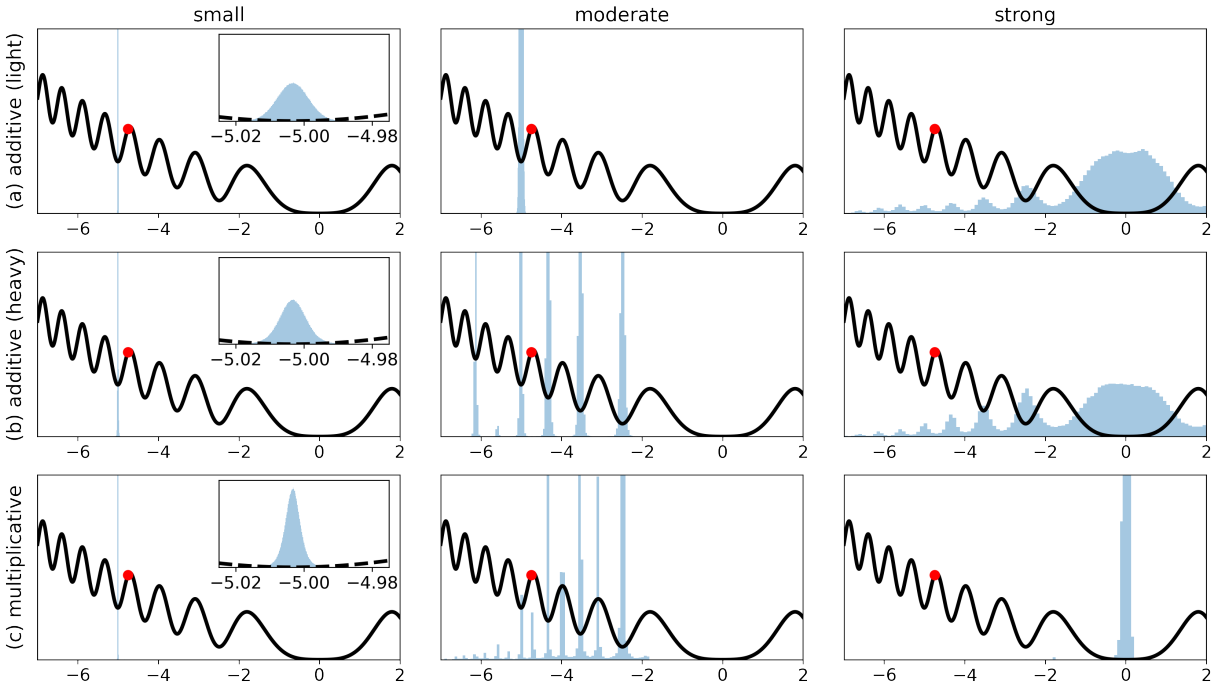

Figure 1: Histograms (blue) of 10^6 iterations of GD with combinations of small (left), moderate (center), and strong (right) versus light additive (a), heavy additive (b), and multiplicative noise (c), applied to a non-convex objective (black). (Red) Initial starting location for the optimization.

This same basin hopping behaviour can be observed heuristically in SGD for higher-dimensional models with the aid of principal component analysis (PCA), although jumps become much more frequent. To see this, we consider fitting a two-layer neural network with 16 hidden units for classification of the Musk data set (Dietterich et al., 1997) (168 attributes; 6598 instances) with cross-entropy loss without regularization and step size $\gamma = 10^{-2}$. Two stochastic optimizers are compared: (a) SGD with a single sample per batch (without replacement), and (b) perturbed GD (Jin et al., 2017), where the state-independent covariance of iterations in (b) is chosen to approximate that of (a) on average. PCA-projected trajectories are presented in Figure 2. SGD frequently exhibits jumps between basins, a feature not shared by perturbed GD with only additive noise.

Finally, to illustrate the effect of depth on tail exponents, we examine stationary behaviour of the magnitude of SGD steps $\|w_{k+1} - w_k\|$ (by triangle inequality, at stationarity, the tail exponent of steps is the same as the weights themselves). We plot histograms of four common wide ResNet architectures trained on CIFAR10 in Figure 3, and provide maximum likelihood estimates of the tail exponents. See Appendix C for implementation details and further numerical analyses of this form supporting claims in §4.

6. Discussion

Heavy tails and generalization. In a recent series of papers (Martin & Mahoney, 2018; 2019; 2020a;b), an empirical (or phenomenological) theory of *heavy-tailed self-regularization* is developed, proposing that sufficiently complex and well-trained deep neural networks exhibit *heavy-tailed mechanistic universality*: the spectral distribution of large weight matrices display a power law whose tail exponent is negatively correlated with generalization performance. If true, examination of this tail exponent provides an indication of model quality, and factors that are positively associated with improved test performance, such as decreased batch size. From random matrix theory, these heavy-tailed spectral distributions may arise due to strong correlations arising in the weight matrices (Bordenave et al., 2011), or heavy-tailed distributions arising in each of the weights over time (Arous & Guionnet, 2008). The reality may be some combination of both; here, in line with Şimşekli et al. (2019; 2020a); Gürbüzbalaban et al. (2020), we have illustrated that the second avenue is at least possible, and relationships between factors of the optimization and the tail exponent agree with the present findings. Theorem 1 implies the general existence of power law tail exponents in the fluctuations of stochastic optimizers, while Şimşekli et al. (2020a) have shown, both theoretically *and* empirically, that *these tail exponents correlate*

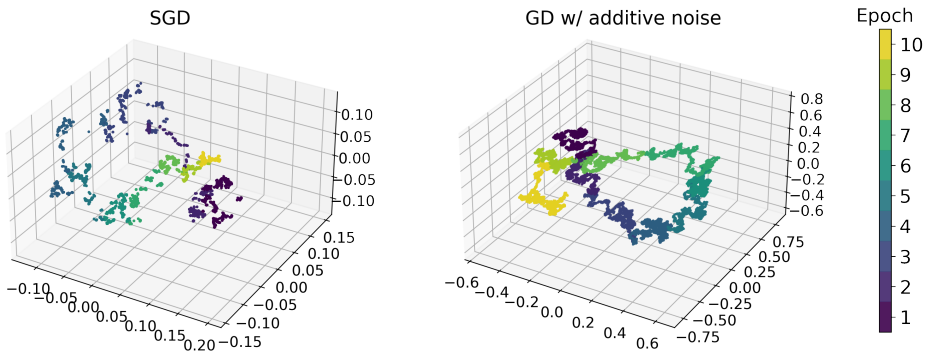

Figure 2: PCA scores for trajectories of SGD over 10 epochs versus perturbed GD with additive noise across the same number of iterations along principle components $2, 3, 4$.

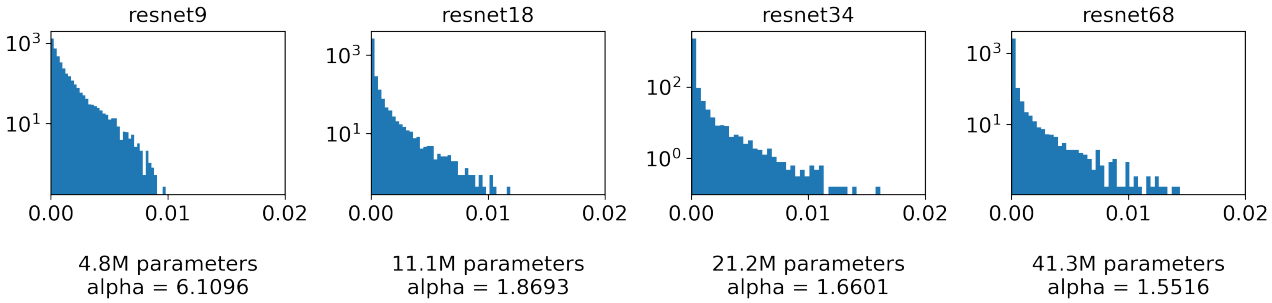

Figure 3: Histograms of $\approx 10^5$ SGD steps $\|w_{k+1} - w_k\|$, number of parameters, and estimate of the corresponding tail exponent α for four ResNet architectures. Note that the tail exponents decrease with depth.

with generalization performance. Proving that heavy-tailed behaviour can arise in spectral distributions of the weights (as opposed to their probability distributions, which we have treated here) is a challenging problem, and remains the subject of future work. Casting the evolution of spectral distributions into the framework of iterated random functions could prove fruitful in this respect.

Lazy training, implicit renewal theory, and the catapult phase. Theory concerning random linear recurrence relations can be extended directly to SGD applied to objective functions whose gradients are "almost" linear, such as those seen in lazy training (Chizat et al., 2019), due to the implicit renewal theory of Goldie (1991). However, it seems unlikely that multilayer networks should exhibit the same tail exponent between each layer, where the conditions of Breiman's lemma (see Appendix A) break down. Significant discrepancies between tail exponents across layers were observed in other appearances of heavy-tailed noise (Martin & Mahoney, 2020b; Şimşekli et al., 2019). From the point of view of tail exponents, it appears that most practical finite-width deep neural networks do not exhibit lazy training. This observation agrees with the poor relative generalization performance exhibited by linearized neural networks (Chizat et al., 2019; Oymak et al., 2019). Indeed, recent efforts have referred to hyperparameter regimes where stochastic optimization differs from (exploitation-focused) lazy training as the *catapult phase* (Lewkowycz et al., 2020). Our analysis suggests that the underlying mechanism for this phenomenon could be attributed to multiplicative noise effects.

Continuous-time models. Probabilistic analyses of stochastic optimization often consider continuous-time approximations, e.g., for constant step size, a stochastic differential equation of the form (Mandt et al., 2016; Orvieto & Lucchi, 2019; Fontaine et al., 2020)

$$\mathrm{d}W_t = -\gamma \nabla f(W_t)\mathrm{d}t + g(W_t)\mathrm{d}X_t, \tag{8}$$

where $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. The most common of these are Langevin models where X is Brownian motion (white noise), although heavy-tailed noise has also been considered (Şimşekli et al., 2019; 2020b). In the case where g is diagonal, as a corollary of (Rubin et al., 2014; Ma et al., 2015), we may determine conditions for heavy-tailed stationary behaviour. Letting $r(w) = \|g(w)^{-2}\nabla f(w)\|$, in Table 2 we present a continuous time

analogue of Table 1 with examples of choices of g that result in each regime when f is quadratic. Langevin algorithms commonly seen in the literature (Mandt et al., 2016; Orvieto & Lucchi, 2019) assume constant or bounded volatility (g), resulting in light-tailed stationary distributions for L^2-regularized loss functions. However, even a minute presence of multiplicative noise (unbounded g) can yield a heavy-tailed stationary distribution (Biró & Jakovác, 2005). Fortunately, more recent analyses are allowing for the presence of multiplicative noise in their approximations (Cheng et al., 2020; Fontaine et al., 2020). Based on our results, it seems clear that this trend is critical for adequate investigation of stochastic optimizers using the continuous-time approach. On the other hand, Şimşekli et al. (2019; 2020b) invoked the generalized central limit theorem (CLT) to propose a continuous-time model with heavy-tailed (additive) Lèvy noise, in response to the observation that stochastic gradient noise is frequently heavy-tailed. In their model, the heaviness of the noise does not change throughout the optimization. Furthermore, (Panigrahi et al., 2019) observed that stochastic gradient noise is typically Gaussian for large batch sizes, which is incompatible with generalized CLT. We have illustrated how multiplicative noise also yields heavy-tailed fluctuations. Alternatively, heavy tails in the stochastic gradient noise for *fixed* weights could be log-normal instead, which are known to arise in deep neural networks (Hanin & Nica, 2019). A notable consequence of the Lèvy noise model is that exit times for basin hopping are essentially independent of basin height, and dependent instead on basin *width*, which is commonly observed with SGD (Xing et al., 2018). In the absence of heavy-tailed additive noise, we observed similar behaviour in our experiments for large multiplicative noise, although precise theoretical treatment remains the subject of future work.

Stochastic gradient MCMC. As mentioned in Mandt et al. (2017), the differences between a stochastic optimizer and its corresponding SG-MCMC variant lie in a step-size schedule, and possibly the inclusion of further additive noise. Therefore, the same principles discussed here also apply to stochastic gradient MCMC (SG-MCMC) algorithms (Ma et al., 2015). The presence of multiplicative noise due to data subsampling suggests that the actual stationary distribution of SG-MCMC can be heavy-tailed, regardless of the chosen target distribution.

Geometric properties of multiplicative noise. It has been suggested that increased noise in SGD acts as a form of convolution, smoothing the effective landscape (Kleinberg et al., 2018). This appears to be partially true. As seen in Figure 1, smoothing behaviour is common for additive noise, and reduces resolution in the troughs. Multiplicative noise, which SGD also exhibits, has a different effect. As seen in Rubin et al. (2014), in the continuous-time setting, multiplicative noise equates to conducting additive noise on a modified (via Lamperti transform) loss landscape. There is also a natural interpretation of multiplicative noise through choice of geometry in Riemannian Langevin diffusion (Girolami & Calderhead, 2011; Ma et al., 2015). Under either interpretation, it appears multiplicative noise shrinks the width of peaks and widens troughs in the effective loss landscape, potentially negating some of the undesirable side effects of additive noise.

7. Conclusion

A theoretical analysis on the relationship between multiplicative noise and heavy-tailed fluctuations in stochastic optimization algorithms has been conducted. We propose that viewing stochastic optimizers through the lens of Markov process theory and examining stationary behaviour is key to understanding exploratory behaviour on non-convex landscapes in the initial phase of training. Our results suggest that heavy-tailed fluctuations may be more common than previous analyses have suggested, and they further the hypothesis that such fluctuations are correlated to improved generalization performance. From this viewpoint, we maintain that multiplicative noise should not be overlooked in future analyses on the subject. We have made efforts to develop as general an analysis as possible, however, there remain some limitations. The time-homogeneous Markov chain model (2) is not sensible when data is cycled through in the same order between epochs. Here, a dynamical systems point-of-view is more sensible, with stationary behaviour examined as limit cycles (Chaudhari & Soatto, 2018). It is likely that tighter bounds on the tail exponents could be achieved through more fine-grained analyses on specific model classes. In particular, tighter estimates illuminate more precise relationships between tail exponents and specific factors, most notably dimension.

Acknowledgments. We would like to acknowledge DARPA, NSF, and ONR for providing partial support of this work.

References

Alsmeyer, G. On the Harris recurrence of iterated random Lipschitz functions and related convergence rate results. *Journal of Theoretical Probability*, 16(1):217–247, 2003.

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

Alstott, J. and Bullmore, D. P. powerlaw: a Python package

for analysis of heavy-tailed distributions. *PLOS One*, 9 (1), 2014.

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

Babichev, D. and Bach, F. Constant step size stochastic gradient descent for probabilistic modeling. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, 2018.

Balles, L., Romero, J., and Hennig, P. Coupling adaptive batch sizes with learning rates. *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, 2017.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. *Regular variation*, volume 27. Cambridge University Press, 1989.

Biró, T. S. and Jakovác, A. Power-law tails from multiplicative noise. *Physical Review Letters*, 94(13):132302, 2005.

Blum, J. R. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.

Bordenave, C., Caputo, P., Chafai, D., et al. Spectrum of large random reversible Markov chains: heavy-tailed weights on the complete graph. *Annals of Probability*, 39(4):1544–1590, 2011.

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 *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of Adam-type algorithms for non-convex optimization. *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, 2019.

Cheng, X., Yin, D., Bartlett, P., and Jordan, M. Stochastic gradient and Langevin processes. In *International Conference on Machine Learning*, pp. 1810–1819. PMLR, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

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

Deutsch, J. M. Generic behavior in linear systems with multiplicative noise. *Physical Review E*, 48(6):R4179, 1993.

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

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pp. 1067–1077, 2017.

Fontaine, X., De Bortoli, V., and Durmus, A. Continuous and Discrete-Time Analysis of Stochastic Gradient Descent for Convex and Non-Convex Functions. *arXiv preprint arXiv:2004.04193*, 2020.

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

Frisch, U. and Sornette, D. Extreme deviations and applications. *Journal de Physique I*, 7(9):1155–1171, 1997.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *JMLR: Workshop and Conference Proceedings*, volume 40, pp. 797–842, 2015.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

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

Goldie, C. M. and Grübel, R. Perpetuities with thin tails. *Advances in Applied Probability*, 28(2):463–480, 1996.

Golmant, N., Vemuri, N., Yao, Z., Feinberg, V., Gholami, A., Rothauge, K., Mahoney, M. W., and Gonzalez, J. On the computational inefficiency of large batch sizes for stochastic gradient descent. *arXiv preprint arXiv:1811.12941*, 2018.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Granas, A. and Dugundji, J. *Fixed point theory*. Springer Science & Business Media, 2013.

Grey, D. R. Regular variation in the tail behaviour of solutions of random difference equations. *The Annals of Applied Probability*, pp. 169–183, 1994.

Grincevičius, A. K. One limit distribution for a random walk on the line. *Lithuanian Mathematical Journal*, 15 (4):580–589, 1975.

Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. The Heavy-Tail Phenomenon in SGD. *arXiv preprint arXiv:2006.04740*, 2020.

Hanin, B. and Nica, M. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, pp. 1–36, 2019.

Henderson, D., Jacobson, S. H., and Johnson, A. W. The theory and practice of simulated annealing. In *Handbook of Metaheuristics*, pp. 287–319. Springer, 2003.

Holland, M. J. Robust descent using smoothed multiplicative noise. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, 2019.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. *Proceedings of Artificial Neural Networks and Machine Learning (ICANN 2018)*, 2018.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pp. 1724–1732. JMLR.org, 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. *Proceedings of the 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.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

Kleinberg, R., Li, Y., and Yuan, Y. An alternative view: When does SGD escape local minima? *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Knuth, D. E. Big omicron and big omega and big theta. *ACM Sigact News*, 8(2):18–24, 1976.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Kroese, D. P., Taimre, T., and Botev, Z. I. *Handbook of Monte Carlo methods*, volume 706. John Wiley & Sons, 2013.

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, M., Yumer, E., and Ramanan, D. Budgeted Training: Rethinking Deep Neural Network Training Under Resource Constraints. *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Liu, S., Papailiopoulos, D., and Achlioptas, D. Bad global minima exist and SGD can reach them. *arXiv preprint arXiv:1906.02613*, 2019.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

Mandt, S., Hoffman, M., and Blei, D. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pp. 354–363, 2016.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Martin, C. H. and Mahoney, M. W. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

Martin, C. H. and Mahoney, M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

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

Martin, C. H. and Mahoney, M. W. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 505–513. SIAM, 2020a.

Martin, C. H. and Mahoney, M. W. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *arXiv preprint arXiv:2002.06716*, 2020b.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

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

Nagaraj, D., Netrapalli, P., and Jain, P. SGD without replacement: Sharper rates for general smooth convex functions. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

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

Orvieto, A. and Lucchi, A. Continuous-time models for stochastic optimization algorithms. In *Advances in Neural Information Processing Systems*, pp. 12589–12601, 2019.

Oymak, S., Fabian, Z., Li, M., and Soltanolkotabi, M. Generalization, Adaptation and Low-Rank Representation in Neural Networks. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 581–585. IEEE, 2019.

Pál, K. F. Hysteretic optimization, faster and simpler. *Physica A: Statistical Mechanics and its Applications*, 360 (2):525–533, 2006.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-gaussianity of stochastic gradient noise. *Science meets Engineering of Deep Learning (SEDL) workshop, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Pennington, J. and Bahri, Y. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pp. 2798–2806. PMLR, 2017.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Roosta-Khorasani, F. and Mahoney, M. W. Sub-sampled Newton methods II: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016.

Rubin, K. J., Pruessner, G., and Pavliotis, G. A. Mapping multiplicative to additive noise. *Journal of Physics A: Mathematical and Theoretical*, 47(19):195001, 2014.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Şimşekli, U., Sagun, L., and Gürbüzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. *arXiv preprint arXiv:2006.09313*, 2020a.

Şimşekli, U., Zhu, L., Teh, Y. W., and Gürbüzbalaban, M. Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise. *arXiv preprint arXiv:2002.05685*, 2020b.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Sornette, D. and Cont, R. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *Journal de Physique I*, 7(3):431–444, 1997.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Volpe, G. and Wehr, J. Effective drifts in dynamical systems with multiplicative noise: a review of recent progress. *Reports on Progress in Physics*, 79(5):053901, 2016.

Wang, J. and Perez, L. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, pp. 11, 2017.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V., and Zhu, Z. On the Noisy Gradient Descent that Generalizes as SGD. *arXiv preprint arXiv:1906.07405*, 2020.

Xing, C., Arpit, D., Tsirigotis, C., and Bengio, Y. A Walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

Yao, Z., Gholami, A., Lei, Q., Keutzer, K., and Mahoney, M. W. Hessian-based analysis of large batch training and robustness to adversaries. In *Advances in Neural Information Processing Systems*, pp. 4949–4959, 2018.

Zhang, G., Li, L., Nado, Z., Martens, J., Sachdeva, S., Dahl, G., Shallue, C., and Grosse, R. B. Which algorithmic choices matter at which batch sizes? Insights from a noisy quadratic model. In *Advances in Neural Information Processing Systems*, pp. 8194–8205, 2019a.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why ADAM beats SGD for attention models. *arXiv preprint arXiv:1912.03194*, 2019b.

Zhang, Z., Zhang, Y., and Li, Z. Removing the feature correlation effect of multiplicative noise. In *Advances in Neural Information Processing Systems*, pp. 627–636, 2018.

Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.