Supplementary material for "Multiplicative Noise and Heavy Tails in Stochastic Optimization"

A. Random linear recurrence relations

Here, we shall discuss existing theory concerning the random linear recurrence relation $W_{k+1} = A_k W_k + B_k$ that arises in (5). Because (A_k, B_k) for each $k = 0, 1, 2, \dots$ is independent and identically distributed, we let $(A, B) = (A_0, B_0)$, noting that $(A, B) \overset{\mathcal{D}}{=} (A_k, B_k)$ for all k. First, we state conditions under which (5) yields an *ergodic Markov chain*. For clarity, we recall the definition of ergodicity in Markov chains (Meyn & Tweedie, 2012, Theorem 13.0.1 and 16.2.1).

Definition 1 (Ergodicity and Geometric Ergodicity)**.** A Markov chain $\{W_k\}_{k=0}^{\infty}$ on $\mathbb{R}^d$ is *ergodic* if there exists a unique invariant probability measure (a *stationary distribution*) π such that for any $x \in \mathbb{R}^d$,

$$\sup_E |\mathbb{P}(W_k \in E | W_0 = x) - \pi(E)| \to 0, \qquad \text{as } k \to \infty,$$

where the supremum is taken over all Borel subsets of $\mathbb{R}^d$. The Markov chain is also *geometrically ergodic* if there exist constants $R > 0$ and $\rho < 1$ such that for any $x \in \mathbb{R}^d$ and Borel set E,

$$|\mathbb{P}(W_k \in E | W_0 = x) - \pi(E)| \leqslant R\rho^k, \qquad \text{for all } k = 0, 1, \dots$$

The following lemma combines Buraczewski et al. (2016, Theorem 4.1.4 and Proposition 4.2.1) and implies Lemma 1.

Lemma 4. *Suppose that A and B are non-deterministic and both $\log^+ \|A\|$ and $\log^+ \|B\|$ are integrable. Then if $\mathbb{E} \log \|A\| < 0$, the Markov chain (5) has a unique stationary distribution. If also either A or B is non-atomic, then the Markov chain (5) is ergodic.*

The intuition behind the presence of heavy-tailed behaviour is easily derived from the Breiman lemma concerning regularly varying random vectors. Recall that a random vector X is regularly varying if there exists a measure μ_X with zero mass at infinity such that

$$\lim_{x \to \infty} \frac{\mathbb{P}(x^{-1} X \in B)}{\mathbb{P}(\|X\| > x)} = \mu_X(B), \quad \text{for any set } B \text{ satisfying } \mu(\partial B) = 0. \tag{9}$$

By Karamata's characterization theorem (Bingham et al., 1989, Theorem 1.4.1), for any regularly varying random vector X, there exists an $\alpha > 0$ such that $x^\alpha \mathbb{P}(x^{-1} X \in \cdot)$ converges as $x \to \infty$ to a non-null measure. In particular, $\|X\|$ and $|\langle u, X \rangle|$ for every $u \in \mathbb{R}^d$ obey a power law with tail exponent α subject to slowly varying functions[1] L, L_u:

$$\mathbb{P}(\|X\| > x) \sim L(x) x^{-\alpha}, \qquad \mathbb{P}(|\langle u, X \rangle| > x) \sim L_u(x) x^{-\alpha}. \tag{10}$$

A random vector X satisfying (9) and (10) is said to be *regularly varying with index α* (abbreviated α-RV). Key to the universality of power laws is a certain closure property of α-RV random variables. This may be viewed analogously to the closure of Gaussian distributions under linear operations and its importance in the central limit theorem. The following is due to Buraczewski et al. (2016, Lemma C.3.1).

Lemma 5 (BREIMAN'S LEMMA)**.** *Let X be an α-RV random vector, A a random matrix such that[2] $\mathbb{E}\|A\|^{\alpha+\epsilon} < +\infty$, and B a random vector such that $\mathbb{P}(\|B\| > x) = o(\mathbb{P}(\|X\| > x))$ as $x \to \infty$. Then $AX + B$ is α-RV.*

In other words, the index of regular variation is preserved under random linear operations, and so regularly varying random vectors are distributional fixed points of random linear recurrence relations. Conditions for the converse statement are well-known in the literature (Buraczewski et al., 2016). Here, we provide brief expositions of the three primary regimes dictating the tails of any stationary distribution of (5). It is worth noting that other corner cases do exist, including super-heavy tails (see Buraczewski et al. (2016, Section 5.5) for example), but are outside the scope of this paper.

[1] Recall that a function f is slowly varying if $f(tx)/f(x) \to 1$ as $\|x\| \to \infty$, for any $t > 0$.

[2] For example, A could be β-RV with $\beta > \alpha$.

A.1. The Goldie–Grübel (light-tailed) regime

To start, consider the case where neither A nor B are heavy-tailed and the stochastic optimization dynamics are such that W_∞ is light-tailed. In particular, assume that all moments of B are finite. By applying the triangle inequality to (5), one immediately finds

$$\|W_{k+1}\| \leqslant \|A_k\|\|W_k\| + \|B_k\|, \quad \text{and} \quad \|W_{k+1}\|_\alpha \leqslant \|A\|_\alpha \|W_k\|_\alpha + \|B\|_\alpha.$$

Therefore, if $\|A\| \leqslant 1$ almost surely and $\mathbb{P}(\|A\| < 1) > 0$, then for any $\alpha \geqslant 1$, $\|A\|_\alpha < 1$ and so $\|W_k\|_\alpha$ is bounded in k. The Markov chain (5) is clearly ergodic, and the existence of all moments suggests that the limiting distribution W_∞ of W_k cannot satisfy a power law. With significant effort, one can show that more is true: Goldie & Grübel (1996, Theorem 2.1) proved that if B is also light-tailed, then W_∞ is **light-tailed**. To our knowledge, this is the only setting where one can prove that the Markov chain (5) possesses a light-tailed limiting distribution, and it requires contraction (and therefore, consistent linear convergence) *at every step, with probability one*. In the stochastic optimization setting, the Goldie–Grübel regime coincides with optimizers that have purely exploitative (no explorative) behaviour. Should the chain fail to contract even once, we move outside of this regime and enter the territory of heavy-tailed stationary distributions.

A.2. The Kesten–Goldie (heavy-tailed due to intrinsic factors) regime

Next, consider the case where neither A nor B are heavy-tailed, but the stochastic optimization dynamics are such that W_∞ is heavy-tailed. To consider a *lower bound*, recall that the smallest singular value of A, $\sigma_{\min}(A)$, satisfies $\sigma_{\min}(A) = \inf_{\|w\|=1} \|Aw\|$. Therefore, once again from (5),

$$\|W_{k+1}\| \geqslant \sigma_{\min}(A_k)\|W_k\| - \|B_k\|, \quad \text{and} \quad \|W_{k+1}\|_\alpha \geqslant \|\sigma_{\min}(A)\|_\alpha \|W_k\|_\alpha - \|B\|_\alpha.$$

Assuming that the Markov chain (5) is ergodic with limiting distribution W_∞, by the f-norm ergodic theorem (Meyn & Tweedie, 2012, Theorem 14.0.1), $\|W_\infty\|_\alpha$ is finite if and only if $\|W_k\|_\alpha$ is bounded in k for any initial W_0. However, if $\mathbb{P}(\sigma_{\min}(A) > 1) > 0$, then there exists some $\alpha > 1$ such that $\|\sigma_{\min}(A)\|_\alpha > 1$. If $\|B\|_\alpha$ is finite, then $\|W_k\|_\alpha$ is unbounded when $\|W_0\|_\alpha$ is sufficiently large, implying that W_∞ is **heavy-tailed**.

This suggests that the tails of the distribution of $\|W_\infty\|$ are at least as heavy as a power law. To show they are dictated *precisely* by a power law, that is, W_∞ is α-RV for some $\alpha > 0$, is more challenging. The following theorem is a direct corollary of the Kesten's celebrated theorem (Kesten, 1973, Theorem 6), and Goldie's generalizations thereof in Goldie (1991).

Theorem 2 (KESTEN–GOLDIE THEOREM)**.** *Assume the following:*

- *The Markov chain (5) is ergodic with $W_\infty = \lim_{k\to\infty} W_k$ (in distribution).*
- *The distribution of X has absolutely continuous component with respect to Lebesgue density that has support containing the zero matrix, and Y is non-zero with positive probability.*
- *There exists $s > 0$ such that $\mathbb{E}\sigma_{\min}(A)^s = 1$.*
- *A is almost surely invertible and $\mathbb{E}[\|A\|^s \log^+ \|A\|] + \mathbb{E}[\|A\|^s \log^+ \|A^{-1}\|] < \infty$.*
- *$\mathbb{E}\|B\|^s < \infty$.*

Then W_∞ is α-RV for some $0 < \alpha \leqslant s$. Furthermore, α uniquely satisfies $\lim_{k\to\infty} \|A_k \cdots A_0\|_\alpha^{1/k} = 1$.

A.3. The Grincevičius–Grey (heavy-tailed due to extrinsic factors) regime

Finally, consider the case where B is heavy-tailed, in particular, that B is β-RV. If $\|A\|_\beta < 1$, then the arguments seen in the Kesten–Goldie regime can no longer hold, since $\|B\|_\alpha$ would be infinite for any α such that $\|\sigma_{\min}(A)\|_\alpha = 1$. Instead, by Buraczewski et al. (2016, Theorem 4.4.24), a limiting distribution of (5) is necessarily β-RV, provided that $\|A\|_{\beta+\delta}$ is finite for some $\delta > 0$. This was proved in the univariate case by Grincevičius (1975, Theorem 1), later updated by Grey (1994) to include the converse result: if the limiting distribution of (5) is β-RV and $\|A\|_\beta < 1$, then B is β-RV.

Contrary to the Kesten–Goldie regime, here neither A nor the recursion itself play much of a role. The optimization procedure itself is fairly irrelevant: from Breiman's lemma, the distribution of W_k is heavy-tailed after only the first iteration, and the tail exponent remains constant, i.e., W_∞ is heavy-tailed. Therefore, in the Grincevičius–Grey regime, the dynamics of the stochastic optimization are dominated by **extrinsic factors**.

B. The effect of dimension under the Wishart+Wigner model

At full generality, it is impossible to directly assess the effect of increasing dimension on the tail exponent of the stationary distribution. This is true even in the linear case (5), as the spectral distribution of the multiplicative factor A_k (and hence the tail exponent, itself) is not wholly dependent on the dimension. Therefore, to investigate relative dependence on dimension, some assumptions are necessary.

Focusing exclusively on SGD, from (7a), it will suffice to assume the form of the spectral distribution of the Hessian $H = \nabla^2 \ell(w, X)$. For neural network models, one natural approximation of this distribution can be found in Pennington & Bahri (2017). Here, the Hessian decomposes into two pieces: a positive-semidefinite matrix $H_0 = JJ^\top$, where J is a Jacobian matrix; and a matrix of second derivatives H_1. Both J and H_1 are assumed to be comprised of weakly dependent entries, in line with universality laws for random matrices. Letting n denote the "effective size" of the data set (the product of the number of classes with the batch size), and p the dimension of the weights w, the spectral distribution of H_0 is naturally modelled in the large n, p regime by the Marchenko–Pastur distribution:

$$\rho_{\text{MP}}(\lambda) = \left(1 - \frac{n}{p}\right)_+ \delta(\lambda) + \frac{n}{2\pi\lambda\sigma p}\sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)}, \qquad \text{for } \lambda \in [\lambda_-, \lambda_+],$$

where σ^2 is the variance of the entries of J, $\lambda_\pm = \sigma(1 \pm \sqrt{p/n})^2$ and $\delta(\lambda)$ is the Dirac delta distribution at the origin. On the other hand, under the same large n, p regime, the spectral distribution of H_1 can be modelled by the Wigner semicircle law:

$$\rho_{\text{SC}}(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}, \qquad \text{for } |\lambda| \leqslant 2\sigma.$$

The spectral distribution of the sum of these two random matrices can be derived using the $\mathcal{R}$ transform — for further details, see Pennington & Bahri (2017, §4.1).

To simplify matters, we shall focus on the underparameterized setting, where $p < n$, and hence, H_0 is non-singular. Under the Wishart+Wigner model, only the Wishart component involves the dimension p, so we restrict our attention to that. If the step size γ is sufficiently small (this also helps to ensure ergodicity of the SGD Markov chain model), the smallest singular value of $I - \gamma H$ will decrease as the support of the Marchenko–Pastur distribution shifts towards the origin. Keeping all else constant, this occurs when p is increased. Hence, k_Ψ is expected to grow with the dimension, and so the tail exponent should decrease (revealing heavier tails).

Unfortunately, the situation becomes more complicated in the overparameterized setting $p > n$, although we conjecture that the same relationship holds as p/n becomes large. Ignoring the spectrum at zero (which might be justified through a similar trick to that seen in the proof of Lemma 2), once p/n becomes sufficiently large, the spectrum of $I - \gamma H$ will be entirely negative. Beyond this point, k_Ψ grows with the eigenvalue gap in the Marchenko–Pastur distribution, and hence, with the dimension. Nevertheless, a rigorous proof of this relationship remains an open problem, and the subject of future work.

C. Numerical examinations of heavy tails

Power laws are notoriously treacherous to investigate empirically (Clauset et al., 2009), especially in higher dimensions (Panigrahi et al., 2019), and this plays a significant role in our focus on establishing mathematical theory. Nevertheless, due to the mystique surrounding heavy tails and our discussion in §4 concerning the impact of various factors on the tail exponent being predominantly informal, we also recognize the value of empirical confirmation. Here, we shall conduct a few numerical examinations to complement our main discussion. For further empirical analyses concerning non-Gaussian fluctuations in stochastic optimization, we refer to (Şimşekli et al., 2019; Panigrahi et al., 2019).

As a quick illustration, in Figure 4, we contrast tail behaviour in the stationary distributions of the Markov chains induced by optimizers (a) (additive) and (c) (multiplicative) introduced in §5. Three different step sizes are used, with constant $\sigma = 10$. To exacerbate multimodality in the stationary distribution, we consider an objective f with derivative $f'(x) =$

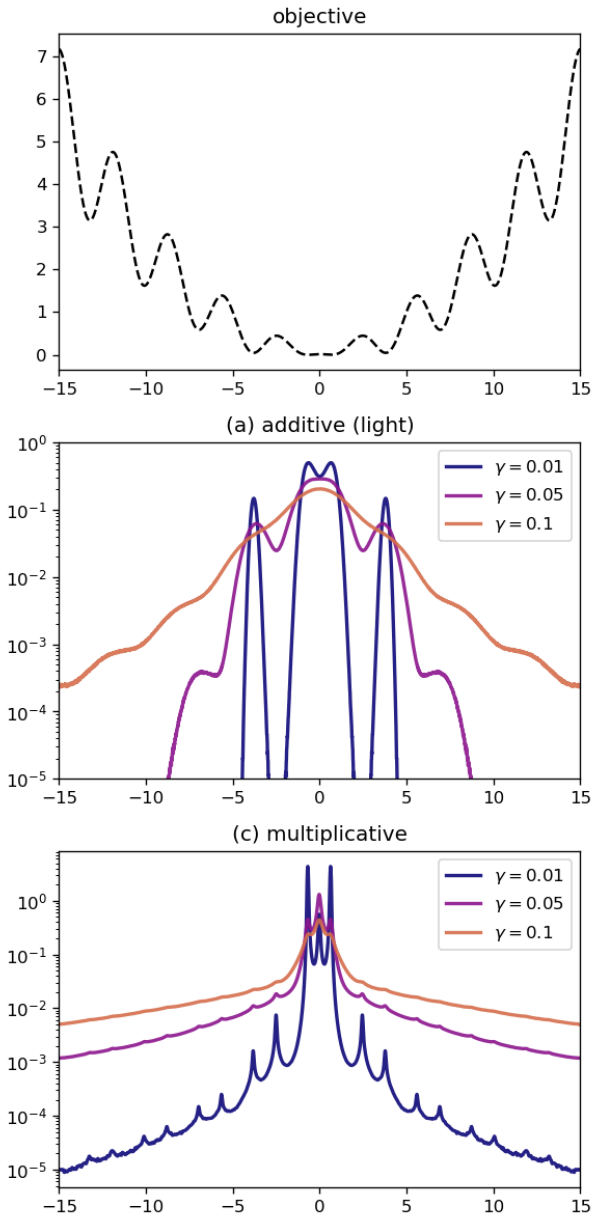

Figure 4: Estimated stationary distributions for optimizers (a) and (c) applied to a non-convex objective f with derivative $f'(x) = x(1 - 4\cos(2x))$, over varying step sizes γ.

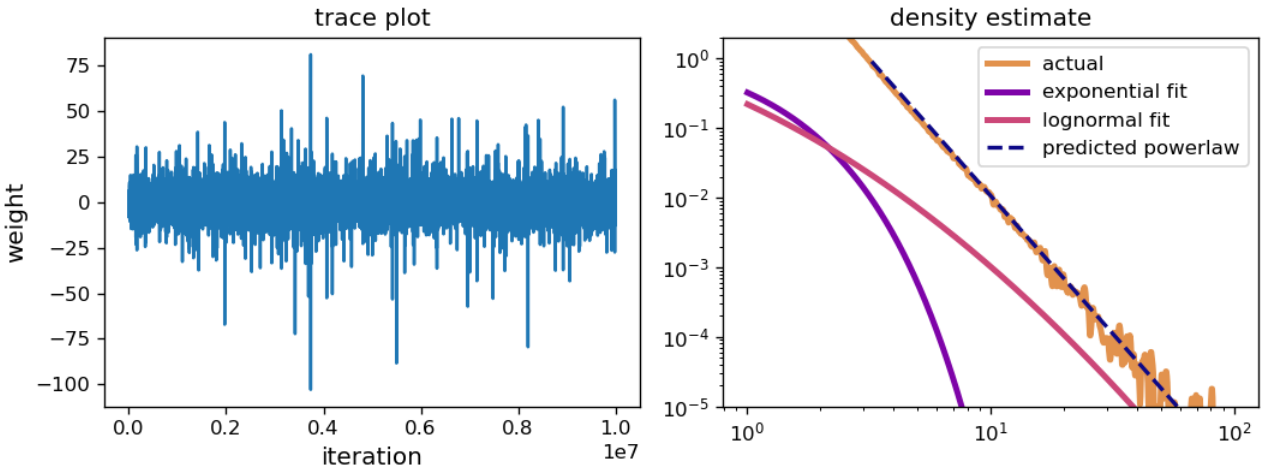

Figure 5: (Left) A trace plot of 10^7 iterations of (11) and (right) a corresponding probability density estimate of their absolute values, with exponential, log-normal fits, and the power law predicted by Theorem 2.

$x(1 - 4\cos(2x))$, visualized in the upper part of Figure 4. Accurately visualizing the stationary distribution, especially its tails, is challenging: to do so, we apply a low bandwidth kernel density estimate to 10^9 steps. As expected, multiplicative noise exhibits slowly decaying heavy tails in contrast to the rapidly decaying Gaussian tails seen with additive light noise. Furthermore, the heaviness of the tails increases with the step size.

To estimate the power law from empirically obtained data, in the sequel, we shall make use of the `powerlaw` Python package (Alstott & Bullmore, 2014), which applies a combination of maximum likelihood estimation and Kolmogorov-Smirnov techniques (see (Clauset et al., 2009)) to fit a Pareto distribution to data. Recall that a Pareto distribution has density $p(t) = \beta t_{\min}^{\beta} t^{-\beta}$ for $t \geqslant t_{\min}$, where $t_{\min}$ is the scale parameter (that is, where the power law in the tail begins), and β is the tail exponent in the density. Note that this β is related to our definition of the tail exponent α by $\alpha = \beta - 1$. Unbiased estimates of this tail exponent α obtained from the `powerlaw` package will be denoted by $\hat{\alpha}$.

C.1. The linear case with SGD

Let us reconsider the simple case discussed in §3 and illustrate power laws arising from SGD on ridge regression. As a particularly simple illustration, first consider the one-dimensional case of (5) with $n = 1$, $\gamma = \frac{1}{2}$, $\lambda = 0$, and standard normal synthetic data. The resulting Markov chain is

$$W_{k+1} = (1 - \tfrac{1}{2}X_k^2)W_k - \tfrac{1}{2}X_k Y_k, \tag{11}$$

where $X_k, Y_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Starting from $W_0 = 0$, Figure 5 shows a trace plot of 10^7 iterations of (11). One can observe the sporadic spikes that are indicative of heavy-tailed fluctuations. Also in Figure 5 is an approximation of the probability density of magnitudes of the iterations. Both exponential and log-normal fits are obtained via maximum likelihood estimation and compared with the power law predicted from Theorem 2 ($\alpha \approx 2.90$). Visually, the power law certainly provides the best fit. Using the Python package `powerlaw`, a Pareto-distribution was fitted to the iterations. The theoretical tail exponent falls within the 95% confidence interval for the estimated tail exponent: $\hat{\alpha} = 2.95 \pm 0.06$. However, even for this simple case where the stationary distribution is known to exhibit a power law and a significant number of samples are available, a likelihood ratio test was found incapable of refuting a Kolmogorov-Smirnov lognormal fit to the tail.

As the dimension increases, the upper bound on the power law from Theorem 2 becomes increasingly less tight. To see this, we conduct least-squares linear regression to the Wine Quality data set (Cortez et al., 2009) (12 attributes; 4898 instances) using vanilla SGD with step size $\gamma = 0.3$, L^2 regularization parameter $\lambda = 4$, and minibatch size $n = 1$. These parameters are so chosen to ensure that the resulting sequence of iterates satisfying (5) is just barely ergodic, and exhibits a profoundly heavy tail. Starting from standard normal W_0, Figure 6 shows a trace plot of 2.5 million iterations, together with an approximation of the probability density of magnitudes of the iterations. A Pareto-distribution fit obtained using the `powerlaw` package is also drawn and can be seen to be an excellent match to the data; the corresponding estimated

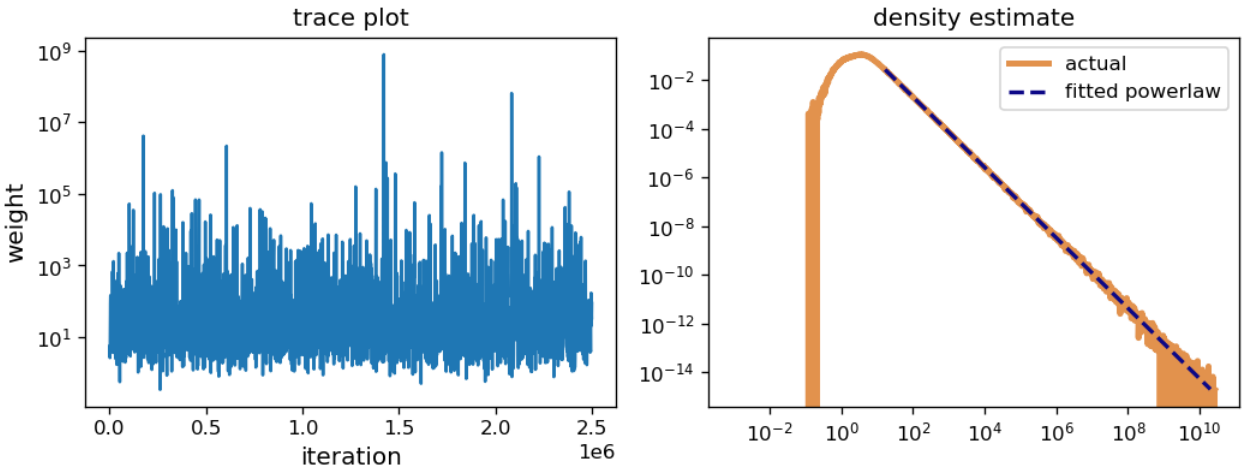

Figure 6: (Left) A trace plot of $2,500,000$ iterations of (5) for the Wine Quality data set and (right) a corresponding probability density estimate of their norms, with fitted Pareto distribution.

tail exponent is $\hat{\alpha} \approx 0.446 \pm 0.008$ to 95% confidence. However, applying Theorem 2 to the chain formed from every 12 iterations reveals a much larger upper bound on the tail exponent: $\alpha \leqslant 22$.

C.2. Factors influencing tail exponents

To help support the claims in §4 concerning the influence of factors on tail exponents, we conducted least-squares regression to the Wine Quality data set (Cortez et al., 2009) (12 attributes; 4898 instances) using a two-layer neural network with four hidden units and ReLU activation function. Our baseline stochastic optimizer is vanilla SGD with a constant step size of $\gamma = 0.025$, minibatch size $n = 1$, and L^2 regularization parameter $\lambda = 10^{-4}$. The effect of changing each of these hyperparameters individually was examined. Three other factors were also considered: (i) the effect of smoothing input data by adding Gaussian noise with standard deviation ϵ; (ii) the effect of adding momentum (and increasing this hyperparameter); and (iii) changing the optimizer between SGD, Adagrad, Adam, and subsampled Newton (SSN).

In each case, we ran the stochastic optimizer for $500n$ epochs (roughly 2.5 million iterations). Instead of directly measuring norms of the weights, we prefer to look at norms of the steps $W_{k+1} - W_k$. There are two reasons for this: (1) unlike steps, the mode of the stationary distribution of the norms of the weights will not be close to zero, incurring a significant challenge to the estimation of tail exponents; and (2) if steps at stationarity are heavy-tailed in the sense of having infinite αth moment, then the stationary distribution of the norms of the weights will have infinite αth moment also. This is due to the triangle inequality: assuming $\{W_k\}_{k=1}^{\infty}$ is ergodic with $W_k \xrightarrow{\mathcal{D}} W_\infty$, $\|W_\infty\|_\alpha \geqslant \frac{1}{2} \limsup_{k\to\infty} \|W_{k+1} - W_k\|_\alpha$. Density estimates for the steps of each run, varying each factor individually, are displayed in Figure 7. Using the `powerlaw` package, tail exponents were estimated in each case, and are presented in Table 1 as 95% confidence intervals. As expected, both increasing step size and decreasing minibatch size can be seen to decrease tail exponents, resulting in heavier tails. Unfortunately, the situation is not as clear for the other factors; from Figure 7, we can see that this is possibly due in part to the unique shapes of the distributions, preventing effective estimates of the scale parameter, upon which the tail exponent (and its confidence intervals) are dependent. Nevertheless, there are a few comments that we can make. Firstly, the inclusion of momentum does not seem to prohibit heavy-tailed behaviour, even though the theory breaks down in these cases. On the other hand, as can be seen in Figure 7, Adam appears to exhibit very light tails compared to other optimizers. Adagrad exhibits heavy-tailed behaviour despite taking smaller steps on average. SSN shows the strongest heavy-tailed behaviour among all the stochastic optimizers considered. Increasing L^2 regularization does increase variance of the steps, but does not appear to make a significant difference to the tails in this test case. Similarly, the effect of adding noise to the data is unclear, although our claim that increasing dispersion of the data (which the addition of large amounts of noise would certainly do) results in heavier-tailed behaviour, is supported by the $\epsilon = 1$ case.

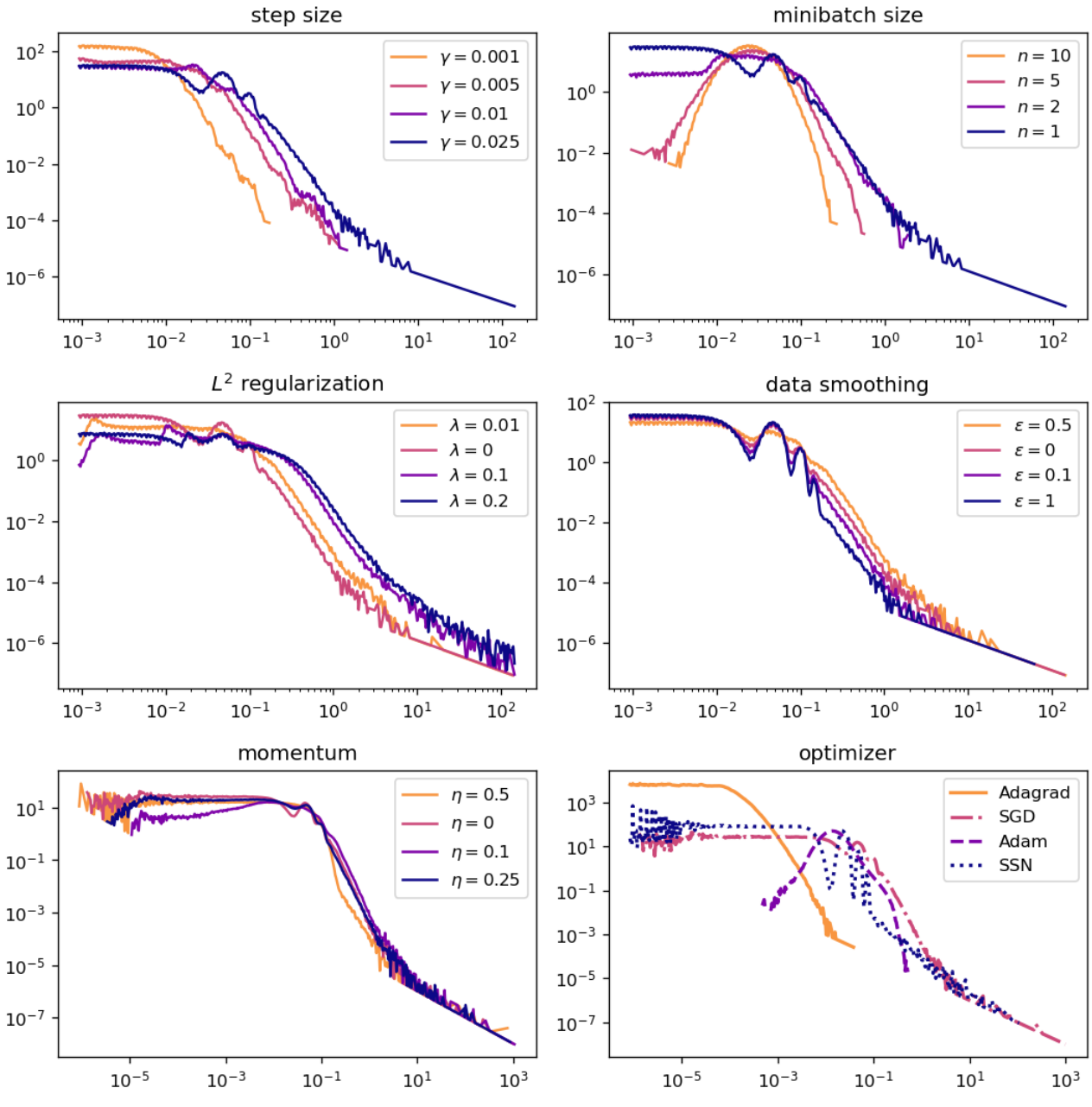

Figure 7: Density estimates of the distributions of norms of steps in roughly 2.5 million stochastic optimization iterations. Baseline hyperparameters are a step size of $\gamma = 0.025$, minibatch size $n = 1$, L^2 regularization parameter $\lambda = 10^{-4}$, no Gaussian perturbations to input data ($\epsilon = 0$), and using SGD with momentum parameter $\eta = 0$. Each plot varies only one of these parameters. For the last plot (bottom right), the Adagrad, Adam, and SSN optimizers are considered in place of SGD, again using the same baseline hyperparameters (and $\beta_1 = 0.9$, $\beta_2 = 0.999$ for Adam). Darker colours indicate smaller estimated tail exponents and heavier tails (from Table 1).

step size		minibatch size		L^2 regularization		data smoothing	
γ	$\hat{\alpha}$	n	$\hat{\alpha}$	λ	$\hat{\alpha}$	ϵ	$\hat{\alpha}$
0.001	4.12 ± 0.04	10	5.99 ± 0.05	10^{-4}	2.97 ± 0.03	0	2.97 ± 0.03
0.005	3.70 ± 0.02	5	4.98 ± 0.07	0.01	3.02 ± 0.02	0.1	2.96 ± 0.05
0.01	3.71 ± 0.04	2	3.62 ± 0.03	0.1	2.77 ± 0.01	0.5	3.05 ± 0.02
0.025	2.97 ± 0.03	1	2.97 ± 0.03	0.2	2.55 ± 0.01	1	2.36 ± 0.13

momentum		optimizer	
η	$\hat{\alpha}$		$\hat{\alpha}$
0.5	4.99 ± 0.03	Adagrad	3.2 ± 0.1
0.25	2.48 ± 0.02	Adam	2.119 ± 0.005
0.1	2.84 ± 0.02	SGD	2.93 ± 0.03
0	2.93 ± 0.03	SSN	0.79 ± 0.04

Table 1: Estimated tail exponents for the distributions of norms of steps in roughly 2.5 million stochastic optimization iterations, varying only one hyperparameter from a baseline step size $\gamma = 0.025$, minibatch size $n = 1$, L^2 regularization parameter $\lambda = 10^{-4}$, no Gaussian perturbations to input data ($\epsilon = 0$), using SGD with momentum parameter $\eta = 0$. The Adagrad, Adam, and SSN optimizers are also considered, using the same baseline hyperparameters (and $\beta_1 = 0.9$, $\beta_2 = 0.999$ for Adam).

C.3. Implementation details for the ResNet experiment

Four wide ResNet architectures were considered without batch normalization: `resnet9`, `resnet18`, `resnet34`, and `resnet68`. Each model was first pretrained on (normalized) CIFAR10 (Krizhevsky & Hinton, 2009) with cross-entropy loss for 200 epochs using SGD with batch size $n = 64$, weight decay $\lambda = 5 \times 10^{-4}$, momentum parameter $\eta = 0.9$, and a cosine annealing step size schedule with a period of 200 epochs, starting with a step size of $\gamma = 0.1$. Minibatches are randomly shuffled at each epoch. After training, five epochs worth of iterations of SGD without momentum are performed with a batch size of $n = 16$, weight decay $\lambda = 5 \times 10^{-4}$, and a constant step size of $\gamma = 5 \times 10^{-4}$. Once again, minibatches are randomly shuffled at each epoch. The norm of the difference in the weights $w_{k+1} - w_k$ between each iteration is recorded. These norms are presented as histograms in each subplot of Figure 3. As in the previous section, the `powerlaw` package was used to obtain estimates of the tail exponents for each of the SGD step distributions.

D. Proofs

Proof of Lemma 2. Observe that $\bar{X}_k = n^{-1} \sum_{i=1}^n X_{ik} X_{ik}^\top$ has full support on the space of square $d \times d$ matrices when $n \geqslant d$, and full support on rank-n $d \times d$ matrices otherwise. In the former (over)determined case, the spectrum of $\bar{X}_k$ has full support on $\mathbb{R}$, implying that $\sigma_{\min}(A_k)$ has full support on $[0, \infty)$, and hence the stationary distribution of (5) decays as a power law by Theorem 2. The latter underdetermined case is not so straightforward: while $\sigma_{\max}(A_k)$ still has full support on $[0, \infty)$, $\sigma_{\min}(A_k) \leqslant |1 - \lambda\gamma|$ almost surely. This is because zero is contained within the spectrum of $\gamma n^{-1} \sum_{i=1}^n X_{ik} X_{ik}^\top$, implying $1 - \lambda\gamma$ is always contained in the spectrum of A_k. This is insufficient to prove the result if λ, γ are sufficiently small. Instead, observe that the Markov chain that arises from taking d steps is also a random linear recurrence relation with the same stationary distribution as (5):

$$W_{k+d} = A_k^{(d)} W_k + B_k^{(d)}, \quad \text{where} \quad A_k^{(d)} = A_{k+d} \cdots A_k, \quad B_k^{(d)} = \sum_{l=0}^{d-1} A_{k+d-1} \cdots A_{k+l+1} B_{k+l}.$$

One may verify that $A_k^{(d)}$ has full support on the space of *all* square $d \times d$ matrices, and hence, $\sigma_{\min}(A_k^{(d)})$ has full support on $[0, \infty)$. Indeed, for any $c > 0$, one could take $X_{1,k+j} = \sqrt{c} e_j$ where e_j is the unit vector with 1 in the j-th coordinate, and $X_{i,k+j} = 0$ for $i > 1$. In this case,

$$A_k^{(d)} = (1 - \lambda\gamma) I - \sum_{j=1}^d \gamma n^{-1} c e_j e_j^\top = (1 - \lambda\gamma - \gamma n^{-1} c) I,$$

implying that $\sigma_{\min}(A_k^{(d)}) = |1 - \lambda\gamma - \gamma n^{-1} c|$. Choosing c appropriately, $\sigma_{\min}(A_k^{(d)})$ can be made to take any value in $[0, \infty)$. Full support for the distribution of $\sigma_{\min}(A_k^{(d)})$ then follows from this special case by continuity of eigenvalues

(Bhatia, 2013, Corollary VI.1.6). Hence, by Theorem 2, (5) will exhibit a stationary distribution with heavy tails. □

Before proceeding to the proofs of results in Section 4, we shall recall a few essential results from the literature on ergodic Markov chains. In the sequel, we let $(S, \|\cdot\|)$ be a separable Banach space and $\Psi : S \times \Omega \to S$ a random function on S. Whenever Ψ is assumed to be almost surely Lipschitz continuous, we let K_Ψ denote the random variable such that $K_\Psi(\omega)$ is the Lipschitz constant of $\Psi(\cdot, \omega)$ for each $\omega \in \Omega$. Our focus is the Markov chain $W_{k+1} = \Psi_k(W_k)$, $k = 0, 1, \ldots$, where each Ψ_k is an independent and identically distributed copy of Ψ. This is known as an *iterated function system* (IFS) corresponding to Ψ. First, we have the following geometric ergodicity result for Lipschitz IFS, which follows by combining Alsmeyer (2003, Theorems 2.1 & 3.2). See also Diaconis & Freedman (1999, Theorem 1.1) for a slightly different formulation.

Theorem 3 (Geometric Ergodicity of Lipschitz IFS)**.** *Assume 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. If K_Ψ and $\|\Psi(w) - w\|$ are integrable for some $w^* \in S$, and $\mathbb{E} \log K_\Psi < 0$, then $\{W_k\}_{k=0}^\infty$ is geometrically ergodic. In particular, there exists $\eta \in (0, 1)$ and an invariant measure π of $\{W_k\}$ such that for $f(w) = 1 + \|w - w^*\|^\eta$,*

$$\sum_{n=0}^{\infty} r^{-n} \sup_{|g| \leqslant f} \left| \mathbb{E}[g(W_k) | W_0 = w] - \int g(w) \mathrm{d}\pi(w) \right| < \infty,$$

for all $w \in S$ and some $r \in (0, 1)$ not depending on w.

Next is Letac's principle, which asserts the intuitive fact that any limit of a continuous IFS must satisfy a distributional fixed point. The version shown here is that seen in Goldie (1991, Theorem 2.1).

Theorem 4 (Letac's Principle)**.** *Assume Ψ is a.s. continuous and that $W_\infty = \lim_{k \to \infty} W_k$ exists almost surely and is independent of W_0. Then the law of W_∞ satisfies the distributional fixed point equation $W_\infty \stackrel{\mathcal{D}}{=} \Psi(W_\infty)$.*

Last is the f-norm ergodic theorem, which provides arguably the most useful tool for us. Here, we present only the parts of the result seen in Meyn & Tweedie (2012, Theorem 14.0.1) that will be useful to us.

Theorem 5 (f-Norm Ergodic Theorem)**.** *Suppose that a Markov chain $\{W_k\}_{k=0}^\infty$ is geometrically ergodic. Let W_∞ be a random variable with law given by the limiting distribution of $\{W_k\}_{k=0}^\infty$. If $f(W_\infty)$ is integrable, then $\mathbb{E}f(W_k) \to \mathbb{E}f(W_\infty) < \infty$. Equivalently, if $\mathbb{E}f(W_k)$ is unbounded, then $f(W_\infty)$ is not integrable.*

For the general case, our strategy is to prove that, for some $\alpha > 0$, $\mathbb{E}|f(W_k)|^\alpha$ diverges as $k \to \infty$. If we assume that W is ergodic with limiting distribution W_∞, then the f-norm ergodic theorem implies that $\mathbb{E}|f(W_k)|^\alpha$ converges if and only if $\mathbb{E}|f(W_\infty)|^\alpha$ is finite, hence the divergence of $\mathbb{E}|f(W_k)|^\alpha$ implies $|f(W_\infty)|$ has infinite α-th moment. In particular, here is the proof of Lemma 3.

Proof of Lemma 3. It suffices to show that $\|f(W_\infty)\|_\beta = +\infty$ for any $\beta > \alpha$, where

$$\alpha \coloneqq \inf_{\epsilon > 0} \frac{1}{\log(1 + \epsilon)} \left| \log \inf_{w \in S} \mathbb{P}\left(\frac{|f(\Psi(w))|}{|f(w)|} > 1 + \epsilon \right) \right|.$$

Let $\epsilon > 0$ be arbitrary. For $w \in S$, let $E_\epsilon(w)$ be the event that $|f(\Psi(w))| > (1 + \epsilon)|f(w)|$. Also, let $p_\epsilon = \inf_{w \in S} \mathbb{P}(E_\epsilon(w)) > 0$ from the hypotheses. Since Ψ is independent of W_k, by laws of conditional expectation, for any $\beta > 0$,

$$\begin{aligned}
\|f(W_{k+1})\|_\beta^\beta &= \mathbb{E}[\mathbb{E}[|f(W_{k+1})|^\beta \,|\, W_k]] \\
&\geqslant \mathbb{E}[\mathbb{P}(E_\epsilon(W_k)|W_k)\mathbb{E}[|f(W_{k+1})|^\beta \,|\, W_k, E_\epsilon(W_k)]] \\
&\geqslant \mathbb{E}[\mathbb{P}(E_\epsilon(W_k)|W_k)(1 + \epsilon)^\alpha |f(W_k)|^\beta] \\
&\geqslant p_\epsilon (1 + \epsilon)^\beta \|f(W_k)\|_\beta^\beta.
\end{aligned}$$

For any $\beta > \alpha$, $p_\epsilon(1 + \epsilon)^\beta > 1$, and hence $\|f(W_k)\|_\beta$ diverges as $k \to \infty$. By the f-norm ergodic theorem, $|f(W_\infty)|^\beta$ cannot be integrable; if it were, then $\|f(W_k)\|_\beta$ would converge to $\|f(W_\infty)\|_\beta < +\infty$. □

The arguments of (Alsmeyer, 2016) almost imply Theorem 1, but are incompatible with the conditions that M_Ψ is non-negative, and k_Ψ can be zero. Instead, more subtle arguments are required; for these, we draw inspiration from (Goldie, 1991; Goldie & Grübel, 1996).

Proof of Theorem 1. From Theorem 3, geometric ergodicity of $\{W_k\}_{k=0}^\infty$ is immediate. Similarly, Letac's principle implies that W_∞ satisfies the distributional fixed point equation $W_\infty \stackrel{\mathcal{D}}{=} \Psi(W_\infty)$.

We shall begin by proving (2). Recall that[3] $\log x = \lim_{s\to 0^+} s^{-1}(x^s - 1)$, and so

$$\lim_{s\to 0^+} \frac{\mathbb{E}K_\Psi^s - 1}{s} = \mathbb{E}\log K_\Psi < 0.$$

Therefore, there exists some $s > 0$ such that $\|K_\Psi\|_s < 1$. Using Hölder's inequality, one finds that $\|K_\Psi\|_{\beta-\epsilon} < 1$ for any $\epsilon > 0$. Likewise, since $K_\Psi < 1$ with positive probability, $k_\Psi < 1$ with positive probability also. Since both $k_\Psi < 1$ and $k_\Psi > 1$ occur with positive probability, there exists $r > 0$ such that $\|k_\Psi\|_r > 1$, and so $\|k_\Psi\|_{\alpha+\epsilon} > 1$ for any $\epsilon > 0$. We now consider similar arguments to the proof of Lemma 3. Since $\{W_k\}_{k=0}^\infty$ is geometrically ergodic, by the f-norm ergodic theorem, for any $\gamma > 0$, $\|W_\infty\|_\gamma$ is finite if and only if $\|W_k\|_\gamma$ is bounded in k. Letting $0 < \epsilon < \delta$, for each $k = 0, 1, \ldots,$

$$\|W_{k+1}\|_{\beta-\epsilon} \leqslant \|K_\Psi\|_{\beta-\epsilon}\|W_k\|_{\beta-\epsilon} + \|K_\Psi\|_{\beta-\epsilon}\|w^*\| + \|\Psi(w^*)\|_{\beta-\epsilon}.$$

Note that $\beta < \alpha$, since $k_\Psi \leqslant K_\Psi$ almost surely. Therefore, $\|\Psi(w^*)\|_{\beta-\epsilon} < \infty$, and since $\|K_\Psi\|_{\beta-\epsilon} < 1$, $\|W_k\|_{\beta-\epsilon}$ is bounded and $\|W_\infty\|_{\beta-\epsilon}$ is finite. By Markov's inequality, $\mathbb{P}(\|W_\infty\| > t) \leqslant \|W_\infty\|_{\beta-\epsilon}^{\beta-\epsilon} t^{-\beta+\epsilon}$ for all $t > 0$. On the other hand, for each $k = 0, 1, \ldots,$

$$\|W_{k+1}\|_{\alpha+\frac{1}{2}\epsilon} \geqslant \|k_\Psi\|_{\alpha+\frac{1}{2}\epsilon}\|W_k\|_{\alpha+\frac{1}{2}\epsilon} - \|k_\Psi\|_{\alpha+\frac{1}{2}\epsilon}\|w^*\| - \|\Psi(w^*)\|_{\alpha+\frac{1}{2}\epsilon},$$

and so $\|W_\infty\|_{\alpha+\frac{1}{2}\epsilon}$ is necessarily infinite. By Fubini's theorem,

$$\|W_\infty\|_{\alpha+\frac{1}{2}\epsilon}^{\alpha+\frac{1}{2}\epsilon} = (\alpha + \tfrac{1}{2}\epsilon)\int_0^\infty t^{\alpha+\frac{1}{2}\epsilon-1}\mathbb{P}(\|W_\infty\| > t)\mathrm{d}t.$$

Therefore, we cannot have that $\limsup_{t\to\infty} t^{\alpha+\epsilon}\mathbb{P}(\|W_\infty\| > t) = 0$, since this would imply

$$\|W_\infty\|_{\alpha+\frac{1}{2}\epsilon}^{\alpha+\frac{1}{2}\epsilon} \leqslant (\alpha + \tfrac{1}{2}\epsilon)\int_0^\infty t^{-1-\frac{1}{2}\epsilon}\mathrm{d}t < +\infty,$$

and hence $\limsup_{t\to\infty} t^{\alpha+\epsilon}\mathbb{P}(\|W_\infty\| > t) > 0$. Repeating these arguments for p in place of $\beta - \epsilon$ and $\alpha + \frac{1}{2}\epsilon$ implies statement (3).

Turning now to a proof of (1), since we have already shown the upper bound, it remains to show that $\mathbb{P}(\|W_\infty\| > t) = \Omega(t^{-\mu})$ for some $\mu > 0$; by Lemma 6 this implies the claimed lower bound. We shall achieve this with the aid of Lemmas 7, 8, and 9. First, since $\mathbb{P}(k_\Psi > 1) > 0$ and $x \mapsto \mathbb{P}(X > x)$ is right-continuous, there exists $\epsilon > 0$ such that $\mathbb{P}(k_\Psi > (1+\epsilon)^2) > 0$. In the sequel, we let $C_{\alpha,\epsilon}$ denote a constant dependent only on α, ϵ, not necessarily the same on each appearance. We may perform the following sequence of steps:

$$\begin{aligned}
\|W_{k+1}\|^\alpha &\geqslant (1+\epsilon)^{-\alpha}(\|W_{k+1}\| + \|\Psi(w^*)\|)^\alpha - C_{\alpha,\epsilon}\|\Psi(w^*)\|^\alpha && (12)\\
&\geqslant (1+\epsilon)^{-\alpha}\|W_{k+1} - \Psi(w^*)\|^\alpha - C_{\alpha,\epsilon}\|\Psi(w^*)\|^\alpha && (13)\\
&\geqslant (1+\epsilon)^{-\alpha}(k_\Psi\|W_k\| - k_\Psi\|w^*\| - M_\Psi)_+^\alpha - C_{\alpha,\epsilon}\|\Psi(w^*)\|^\alpha && (14)\\
&\geqslant (1+\epsilon)^{-2\alpha}k_\Psi^\alpha\|W_k\|^\alpha - C_{\alpha,\epsilon}((1+\epsilon)^{-\alpha}(k_\Psi\|w^*\| + M_\Psi)^\alpha + \|\Psi(w^*)\|^\alpha) && (15)
\end{aligned}$$

Inequality (12) follows from the first inequality of Lemma 7 with $z = \|W_{k+1}\|$ and $y = \|\Psi(w^*)\|$, while (13) follows from reverse triangle inequality. The next inequality (14) involves the assumption (6), followed by an application of the reverse triangle inequality. Finally, (15) follows from an application of the second inequality of Lemma 7 with $x = k_\Psi\|W_k\|$ and

[3] This is readily shown using L'Hôpital's rule.

$y = k_\Psi\|w^*\| + M_\Psi$. To simplify further, let $B_\Psi = k_\Psi\|w^*\| + M_\Psi + \|\Psi(w^*)\|$. Now, bounding the second term of (15) from above, since $(1+\epsilon)^{-\alpha} \leqslant 1$ and $x^\alpha + y^\alpha \leqslant (x+y)^\alpha$ for $\alpha \geqslant 1$, there is

$$\|W_{k+1}\|^\alpha \geqslant (1+\epsilon)^{-2\alpha} k_\Psi^\alpha \|W_k\|^\alpha - C_{\alpha,\epsilon} B_\Psi^\alpha.$$

For $\alpha > 1$ and $k = 0, 1, \ldots$, let $f_k^\alpha(t) = \mathbb{E}[\|W_k\|^\alpha \mathbb{1}_{\|W_k\| \leqslant t}]$. Then

$$f_{k+1}^\alpha(t) \geqslant \mathbb{E}[k_\Psi^\alpha (1+\epsilon)^{-2\alpha} \mathbb{1}_{\{\|W_k\| \leqslant t/K_\Psi - \|w^*\| - \|\Psi(w^*)\|/K_\Psi\}} \|W_k\|^\alpha] - C_{\alpha,\epsilon}\mathbb{E}B_\Psi^\alpha.$$

Let E_c be the event that $\min\{K_\Psi, \|\Psi(w^*)\|\} \leqslant c$, and take $c > 1$ to be some constant sufficiently large so that $\mathbb{P}(k_\Psi > (1+\epsilon)^2 | E_c) > 0$. We may now choose $\alpha > 1$ such that $\mathbb{E}[k_\Psi^\alpha (1+\epsilon)^{-2\alpha} | E_c]\mathbb{P}(E_c) =: a > c$. Doing so, we find that

$$f_{k+1}^\alpha(ct) \geqslant a f_k(t - 1 - \|w^*\|) - C_{\alpha,\epsilon}\mathbb{E}B_\Psi^\alpha, \qquad \text{for any } t > 0,\ k \geqslant 0.$$

This is because, subject to the constraints $K_\Psi \leqslant c$ and $\|\Psi(w^*)\| \leqslant c$, the quantity $(t - \|\Psi(w^*)\|)/K_\Psi - \|w^*\|$ seen in the indicator is bounded below by $t/c - 1 - \|w^*\|$ when $t \geqslant c$, and is negative if $t < c$. By the f-norm ergodic theorem, $f_k^\alpha(t) \to f^\alpha(t)$ pointwise as $k \to \infty$, where $f^\alpha(t) = \mathbb{E}[\|W_\infty\|^\alpha \mathbb{1}_{\|W_\infty\| \leqslant t}]$. Therefore,

$$f^\alpha(ct) \geqslant a f(t - 1 - \|w^*\|) - C_{\alpha,\epsilon}\mathbb{E}B_\Psi^\alpha, \qquad \text{for any } t > 0.$$

By Lemma 8, this implies that there exists some $0 < \gamma < \alpha$ such that $\liminf_{t\to\infty} t^{-\gamma} f^\alpha(t) > 0$, which from Lemma 9, implies that $\mathbb{P}(\|W_\infty\| \geqslant t) = \Omega(t^{-\alpha(\alpha-\gamma)/\gamma})$. □

Lemma 6. *Suppose that $\mathbb{P}(X > x) \geqslant Cx^{-\alpha}$ for all $x \geqslant x_0$. Then there exists $c > 0$ such that $\mathbb{P}(X > x) \geqslant c(1+x)^{-\alpha}$ for all $x \geqslant 0$.*

Proof. Evidently, $\mathbb{P}(X > x) \geqslant C(1+x)^{-\alpha}$ for $x \geqslant x_0$. Treating the $x \leqslant x_0$ setting, let

$$C_0 = \inf_{x \leqslant x_0} \frac{\mathbb{P}(X > x)}{(1+x)^\alpha}.$$

Assume $C_0 = 0$. Since $(1+x)^\alpha$ is bounded for all $x \geqslant 0$, there exists a sequence $\{x_n\}_{n=1}^\infty \subset [0, x_0]$ such that $\mathbb{P}(X > x_n) \to 0$. But since $\{x_n\}_{n=1}^\infty$ is bounded, there exists some subsequence converging to a point $x \leqslant x_0$, which must satisfy $\mathbb{P}(X > x) = 0$, contradicting our hypotheses. Therefore, $C_0 \neq 0$ and the result is shown for $c = \min\{C_0, C\}$. □

Lemma 7. *For any $\alpha > 1$ and $\epsilon > 0$, there exists $C_{\alpha,\epsilon} > 0$ such that for any $x, y, z \geqslant 0$,*

$$z^\alpha \geqslant (1+\epsilon)^{-\alpha}(y+z)^\alpha - C_{\alpha,\epsilon} y^\alpha$$
$$(x-y)_+^\alpha \geqslant (1+\epsilon)^{-\alpha} x^\alpha - C_{\alpha,\epsilon} y^\alpha.$$

Proof. The second of these two inequalities follows from the first by taking $(x-y)_+ = z$. Since the first inequality is trivially the case when $y = 0$, letting $\rho = (1+\epsilon)^\alpha$, it suffices to show that

$$\sup_{z \geqslant 0,\ y > 0} \frac{(y+z)^\alpha - \rho z^\alpha}{\rho y^\alpha} < \infty.$$

Equivalently, parameterizing $z = Ly$ where $L \geqslant 0$, it suffices that $\sup_{L \geqslant 0}[(1+L)^\alpha - \rho L^\alpha] < \infty$, which is evidently the case since $(1 + L^{-1})^\alpha - \rho < 0$ for sufficiently large $L > 0$. □

Lemma 8. *Let $f(t)$ be an unbounded non-decreasing function. If there exists some $a \geqslant c > 1$ and $b, x, t_0 \geqslant 0$ such that for $t \geqslant t_0$,*

$$f(ct) \geqslant a f(t - x) - b, \tag{16}$$

then $\liminf_{t\to\infty} t^{-\gamma} f(t) > 0$ for any $\gamma < \frac{\log a}{\log c}$.

Proof. First, consider the case $x = 0$. Iterating (16), for any $n = 1, 2, \ldots$, and $t > t_0$,

$$f(c^n t) \geqslant \left(f(t) - \frac{b}{a-1} \right) a^n + \frac{b}{a-1}.$$

Let t_1 be sufficiently large so that $t_1 > t_0$ and $f(t) > \frac{b}{a-1}$ for all $t > t_1$. Then for any $\alpha > 0$, letting $[\alpha]$ denote the largest integer less than or equal to α,

$$\begin{aligned} f(c^\alpha t_1) &\geqslant \left(f(c^{\alpha - [\alpha]} t_1) - \frac{b}{a-1} \right) a^{[\alpha]} + \frac{b}{a-1}, \\ &\geqslant \left(f(t_1) - \frac{b}{a-1} \right) a^{\alpha - 1} + \frac{b}{a-1}. \end{aligned}$$

In particular, by choosing $\alpha = \frac{\log t - \log t_1}{\log c}$, $c^\alpha t_1 = t$, and so for any $t > t_1$,

$$f(t) \geqslant a^{-\frac{\log t_1}{\log c} - 1} \left(f(t_1) - \frac{b}{a-1} \right) t^{\frac{\log a}{\log c}} + \frac{b}{a-1}.$$

Now, let t_2 be sufficiently large so that $t_2 \geqslant t_1$ and

$$a^{-\frac{\log t_1}{\log c} - 1} \left(f(t_1) - \frac{b}{a-1} \right) t_2^{\frac{\log a}{\log c}} \geqslant \frac{b}{a-1}.$$

Then for $t \geqslant t_2$ and $C = 2a^{-\frac{\log t_1}{\log c} - 1}(f(t_1) - \frac{b}{a-1})$, $f(t) \geqslant C t^{\frac{\log a}{\log c}}$, and the result follows. Now, suppose that $x > 0$. Let $0 < \epsilon < 1$ and take $t_1 = \max\{t_0, x/\epsilon\}$ so that $t - x \geqslant (1 - \epsilon)t$ for all $t \geqslant t_1$. Then, for all $t \geqslant t_1$,

$$f\left(\frac{c}{1-\epsilon} \cdot t \right) \geqslant a f(t) - b.$$

If we can show the conclusion for the case where $x = 0$, then for $\gamma < \frac{\log a}{\log c}$,

$$\liminf_{t \to \infty} t^{-\gamma \cdot \frac{\log c}{\log c + |\log(1-\epsilon)|}} f(t) > 0.$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

Lemma 9. *Let X be a non-negative random variable and $\alpha > 0$. If there exists some $0 < \gamma < \alpha$ such that*

$$\liminf_{t \to \infty} t^{-\gamma} \mathbb{E}[X^\alpha \mathbb{1}_{X \leqslant t}] > 0,$$

then there is some $C, t_0 > 0$ such that for $t \geqslant t_0$,

$$\mathbb{P}(X \geqslant t) \geqslant C t^{-\frac{\alpha}{\gamma}(\alpha - \gamma)}.$$

Proof. By taking t to be sufficiently large, there exists $c_1 > 0$ such that

$$\begin{aligned} c_1 t^\gamma \leqslant \mathbb{E}[X^\alpha \mathbb{1}_{X \leqslant t}] &= \alpha \int_0^\infty \left(\int_0^\infty \mathbb{1}_{u \leqslant t} \mathrm{d}\mathbb{P}_X(u) \right) \mathbb{1}_{v \leqslant u} v^{\alpha - 1} \mathrm{d}v \\ &= \alpha \int_0^t \mathbb{P}(v \leqslant X \leqslant t) v^{\alpha - 1} \mathrm{d}v \leqslant \alpha \int_0^t \mathbb{P}(X \geqslant v) v^{\alpha - 1} \mathrm{d}v. \end{aligned}$$

On the other hand, observe that for any $b, t > 1$,

$$\begin{aligned} \alpha \int_0^{bt} \mathbb{P}(X \geqslant v) v^{\alpha - 1} \mathrm{d}v &\leqslant \alpha \int_0^t v^{\alpha - 1} \mathrm{d}v + \alpha \mathbb{P}(X \geqslant t) \int_t^{bt} v^{\alpha - 1} \mathrm{d}v \\ &= t^\alpha [1 + \mathbb{P}(X \geqslant t)(b^\alpha - 1)]. \end{aligned}$$

Therefore,

$$\frac{c b^\gamma t^{\gamma - \alpha} - 1}{b^\alpha - 1} \leqslant \mathbb{P}(X \geqslant t).$$

Choosing $b = (2/c) t^{(\alpha - \gamma)/\gamma}$ such that $c b^\gamma t^{\gamma - \alpha} - 1 = 1$, the lemma follows. □

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.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

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

Buraczewski, D., Damek, E., and Mikosch, T. *Stochastic models with power-law tails*. Springer, 2016.

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

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

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

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.

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.

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

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

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

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.

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.