Overparameterization and the Power Law Paradigm

Liam Hodgkinson

Heavy Tails in ML
Structure, Stability, Dynamics
A NeurIPS 2023 Workshop

Power laws are arising all over the place in deep learning, *far more often* than with classical models

Power Law Paradigm

- **Scaling Laws**: behavior of test loss (e.g. LLMs)

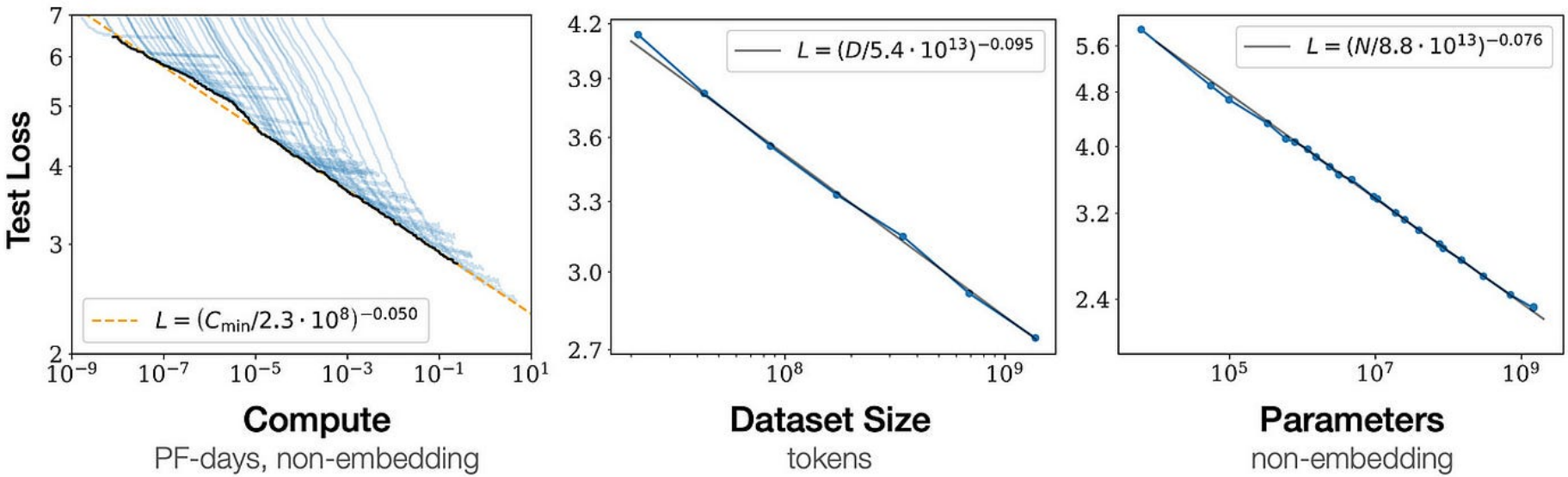


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

- Kaplan, J. et al. (2020). Scaling laws for neural language models.
- Hoffmann, J. et al. (2022). Training compute-optimal large language models.

Power Law Paradigm

- **Scaling Laws**: behavior of test error (e.g. LLMs)
- **Spectral**: power laws appearing in matrix eigenspectra

- The **Hessian** matrix of second derivatives of the loss

Yao, Z., Gholami, A., Keutzer, K., & Mahoney, M. W. (2020). PyHessian: Neural networks through the lens of the Hessian. IEEE International Conference on Big Data (pp. 581-590).

Yao, Z., Gholami, A., Lei, Q., Keutzer, K., & Mahoney, M. W. (2018). Hessian-based analysis of large batch training and robustness to adversaries. Advances in Neural Information Processing Systems, 31.

- The **Gram matrix of the NTK $JJ^\top$**, where J is the Jacobian matrix of first derivatives of the loss

Fan, Z., & Wang, Z. (2020). Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. Advances in Neural Information Processing Systems, 33, 7710-7721.

Karakida, R., Akaho, S., & Amari, S. I. (2021). Pathological spectra of the Fisher information metric and its variants in deep neural networks. Neural Computation, 33(8), 2274-2307.

- Weights of all layers of the network
  - Martin, C. H., Peng, T., & Mahoney, M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications, 12(1), 4122.
  - Martin, C. H., & Mahoney, M. W. (2021). Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. The Journal of Machine Learning Research, 22(1), 7479-7551.
  - Mahoney, M., & Martin, C. (2019). Traditional and heavy tailed self regularization in neural network models. In International Conference on Machine Learning (pp. 4284-4293). PMLR.
- Activation functions of the network
  - Agrawal, K. K., Mondal, A. K., Ghosh, A., & Richards, B. (2022). α-ReQ: Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay. Advances in Neural Information Processing Systems, 35, 17626-17638.

Journal of Machine Learning Research 22 (2021) 1-73 Submitted 4/20; Published 6/21

Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning

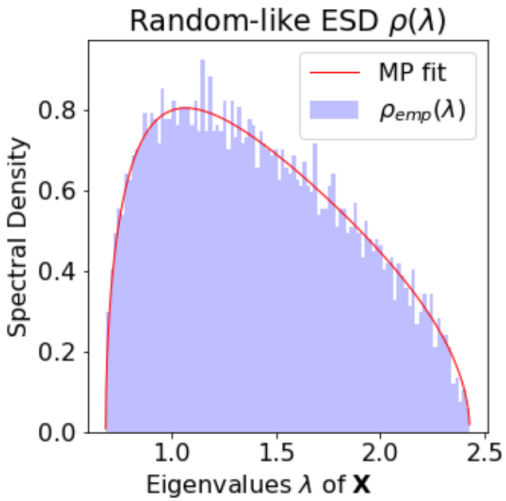
Charles H. Martin CHARLES@CALCULATIONCONSULTING.COM
Calculation Consulting
8 Locksley Ave, 6B
San Francisco, CA 94122

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720

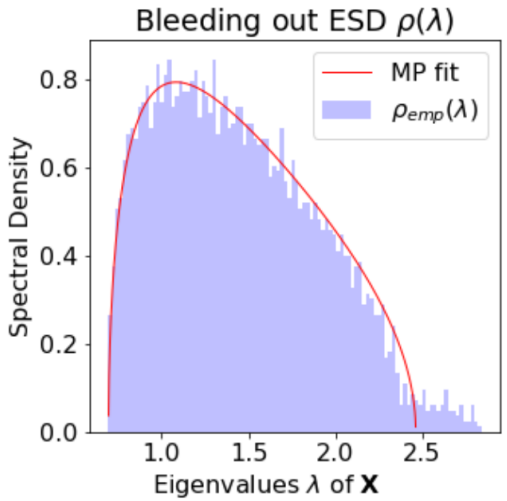
Editor: Ohad Shamir

Abstract

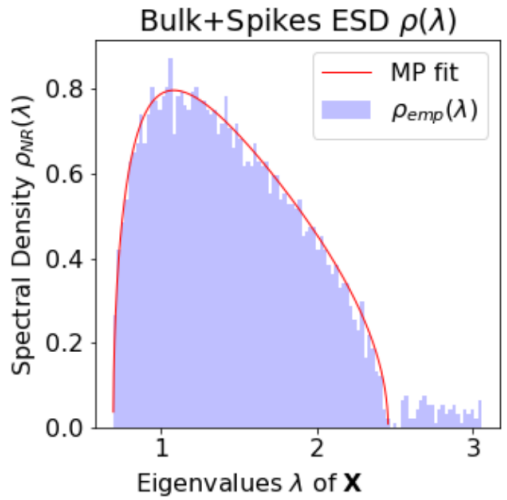
Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neural Networks (DNNs), including both production quality, pre-trained models such as AlexNet and Inception, and smaller models trained from scratch, such as LeNet5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate that the DNN training process itself implicitly implements a form of *Self-Regularization*, implicitly sculpting a more regularized energy or penalty landscape. In particular, the empirical spectral density (ESD) of DNN layer matrices displays signatures of traditionally-regularized statistical models, even



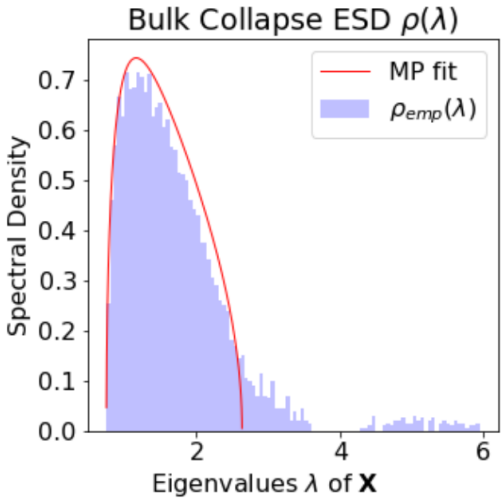
(a) Random-like.



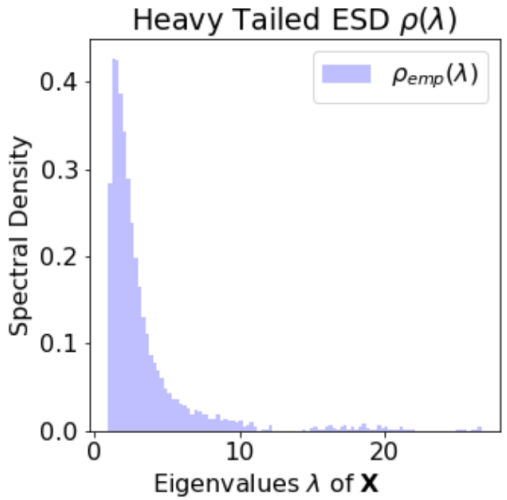
(b) Bleeding-out.



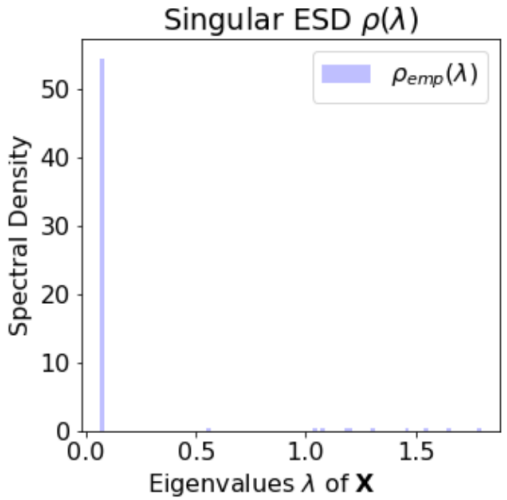
(c) Bulk+Spikes.



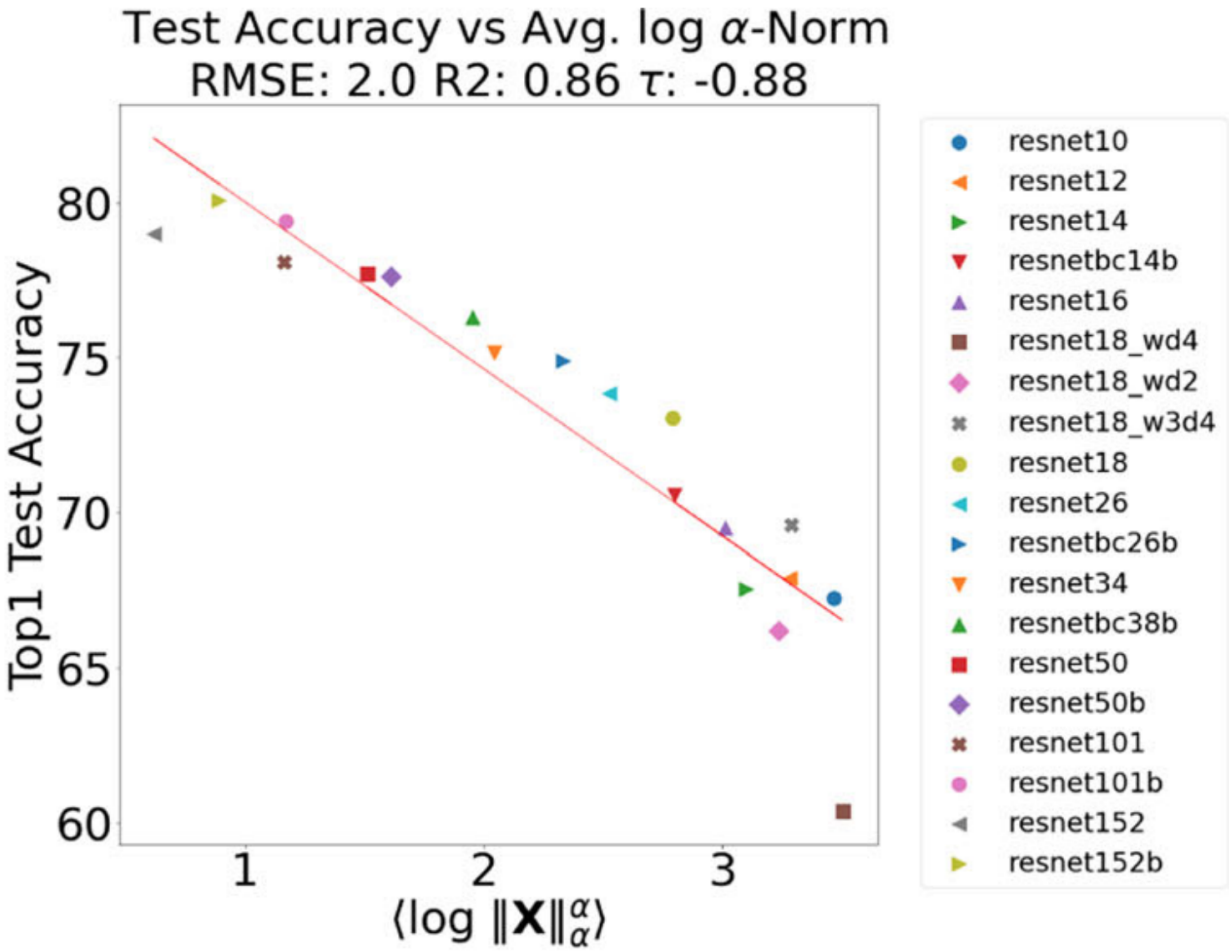
(d) Bulk-decay.



(e) Heavy-Tailed.



(f) Rank-collapse.



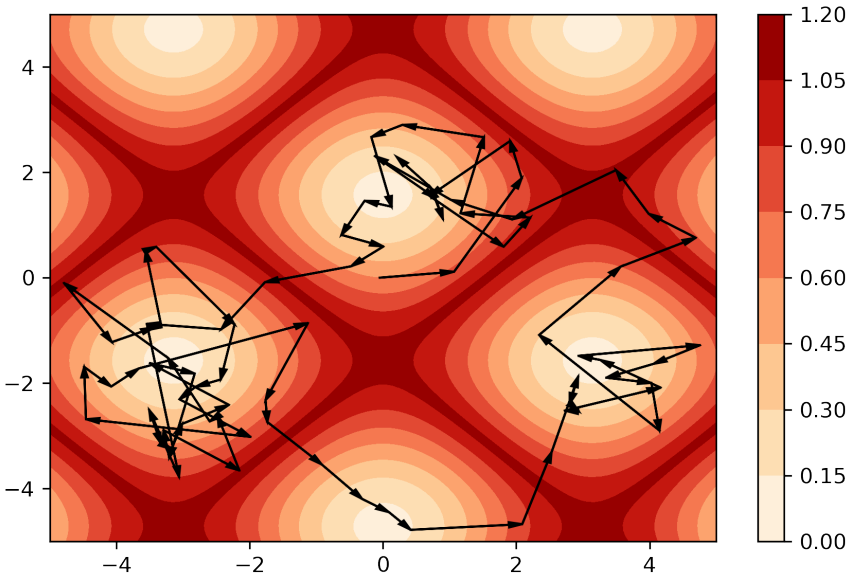
Martin, C. H., Peng, T., & Mahoney, M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications, 12(1), 4122.

Power Law Paradigm

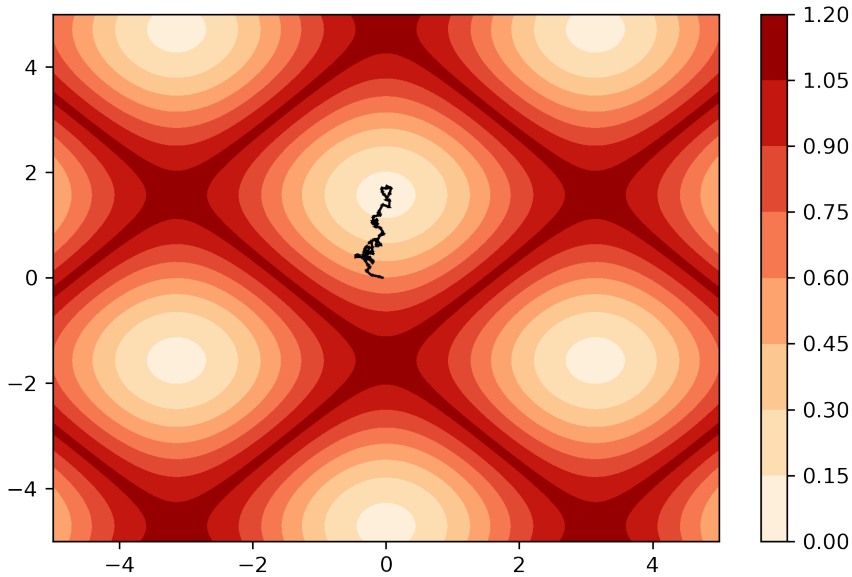
- **Scaling Laws**: behavior of test error (e.g. LLMs)
- **Spectral**: power laws appearing in matrix eigenspectra
- **Dynamics**: heavy-tailed fluctuations during training

Phases of Training

Exploration
large learning rate

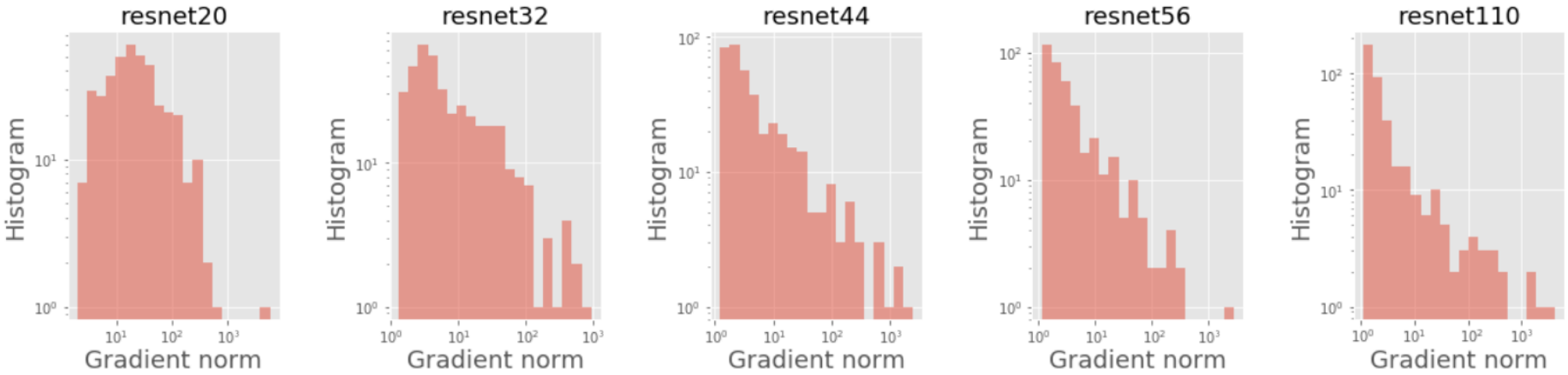


Exploitation
small learning rate



Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., & Gur-Ari, G. (2020). The large learning rate phase of deep learning: the catapult mechanism.

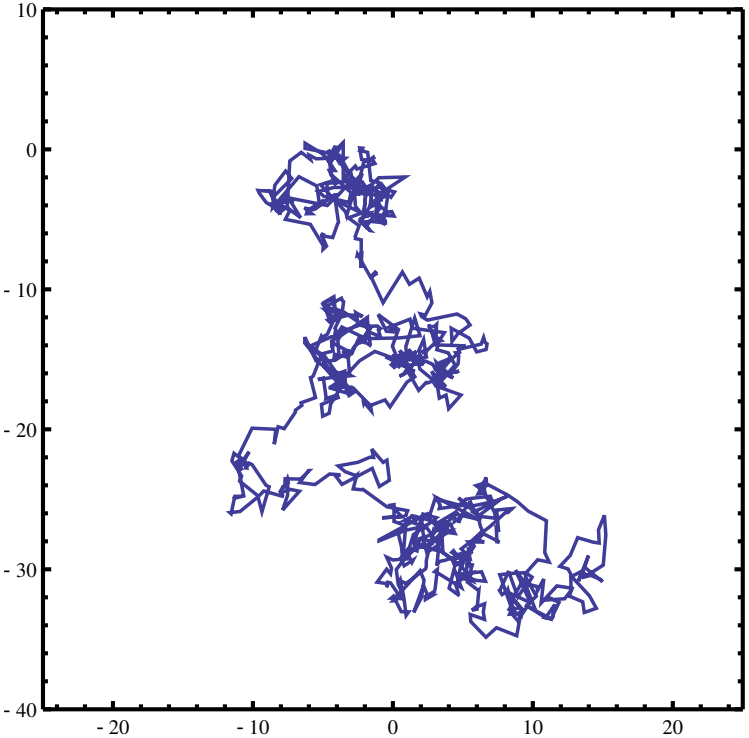
Jumps are heavy-tailed



Simsekli, U., Sagun, L., & Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. International Conference on Machine Learning (pp. 5827-5837).

Brownian motion

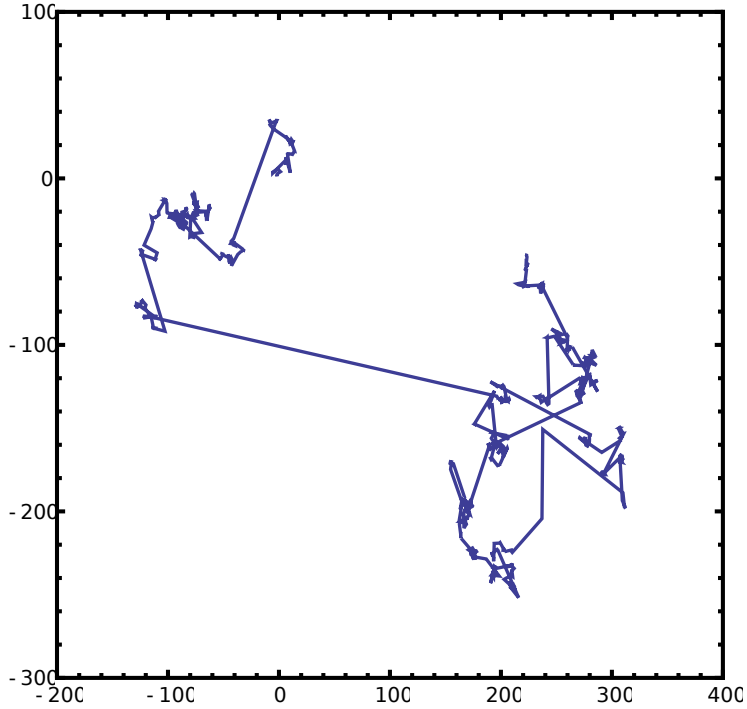
light-tailed



(exponential time to escape basins)

Levy flight

heavy-tailed



(polynomial time to escape basins; Nguyen et al., 2019)

Why might this be happening?

Heavy-tailed data?

Power laws change with lots of other factors too.

Heavy-Tailed Universality

Power Law Paradigm

- **Scaling Laws**: behavior of test error (e.g. LLMs)
- **Spectral**: power laws appearing in matrix eigenspectra
- ***Dynamics***: *heavy-tailed fluctuations during training*

(Gurbuzbalaban et al., 2021)

(Hodgkinson & Mahoney, 2021)

The Heavy-Tail Phenomenon in SGD

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

Abstract

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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 38th International Conference on Machine Learning, PMLR 139, 2021.

(ICML2021): The Kesten Mechanism

Traditionally, stochastic gradient descent

$$w_{k+1} = w_k - \gamma \hat{\nabla} f(w_k)$$

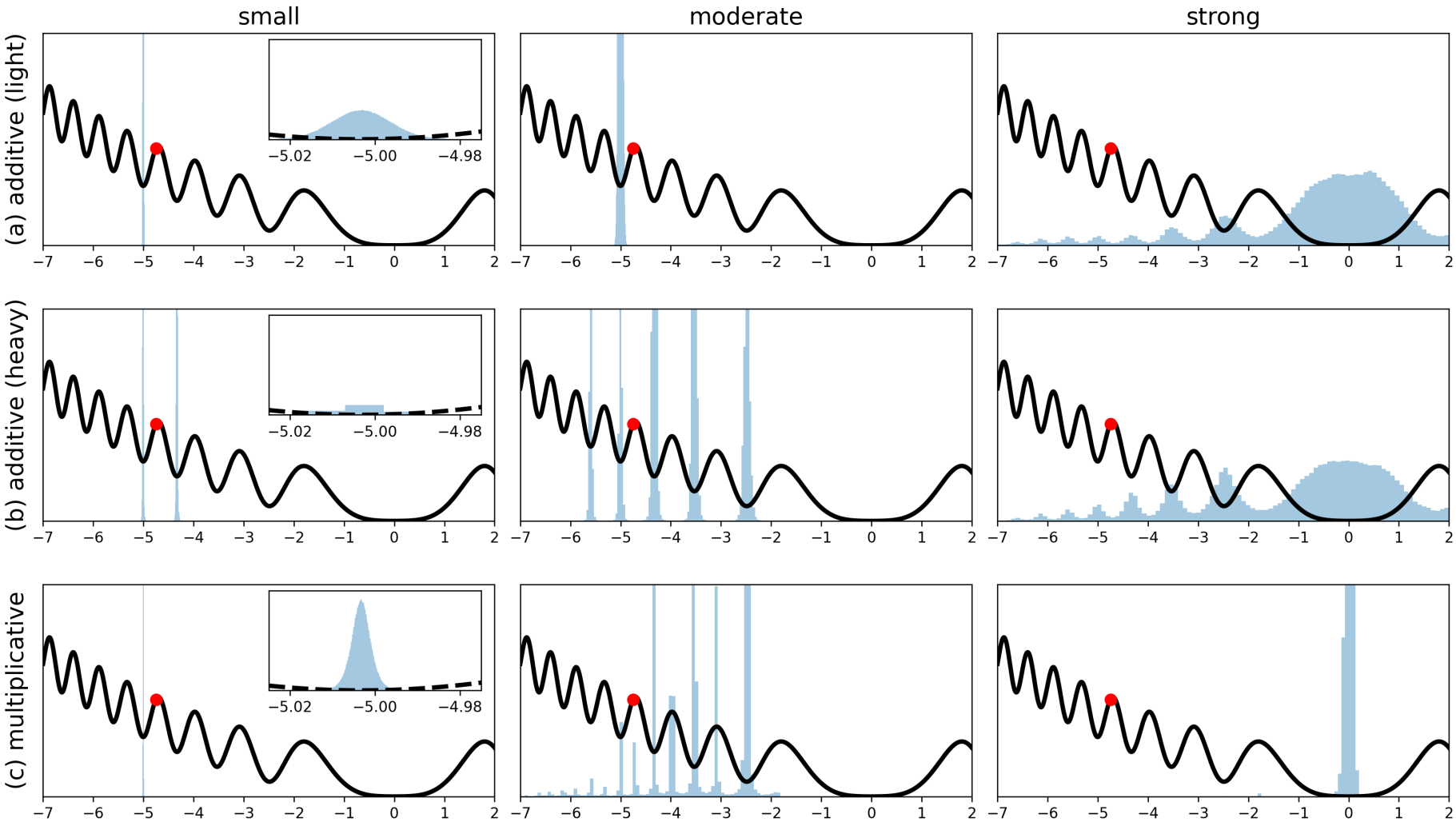
was treated as gradient descent with added noise

$$w_{k+1} = w_k - \gamma \nabla f(w_k) + \epsilon_k.$$

But the covariance is **position-dependent**, so it's closer to multiplicative noise

$$w_{k+1} = w_k - \gamma (I + E_k) \nabla f(w_k) + \epsilon_k$$

(Kesten 1973): If f is linear, w_k has heavy-tailed fluctuations


small
moderate
strong
(a) additive (light)
(b) additive (heavy)
(c) multiplicative
−5.02
−5.00
−4.98
−7
−6
−5
−4
−3
−2
−1
0
1
2

How does this affect generalization performance?

A property at the end of training…

Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks

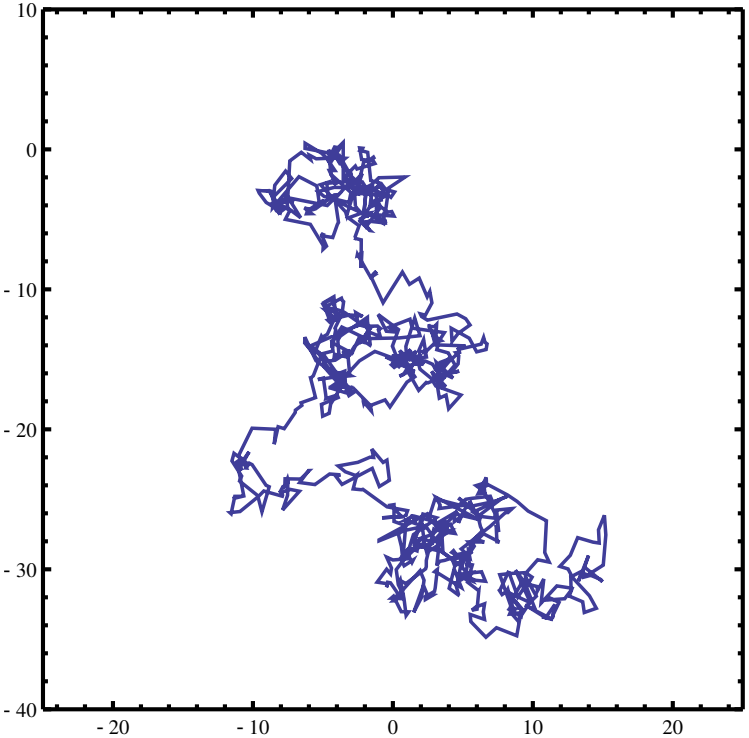
Umut Şimşekli[1,2], Ozan Sener[3], George Deligiannidis[2,4], Murat A. Erdogdu[5,6]
LTCI, Télécom Paris, Institut Polytechnique de Paris[1], University of Oxford[2], Intel Labs[3]
The Alan Turing Institute[4], University of Toronto[5], Vector Institute[6]

Abstract

Despite its success in a wide range of applications, characterizing the generalization properties of stochastic gradient descent (SGD) in non-convex deep learning problems is still an important challenge. While modeling the trajectories of SGD via stochastic differential equations (SDE) under heavy-tailed gradient noise has recently shed light over several peculiar characteristics of SGD, a rigorous treatment of the generalization properties of such SDEs in a learning theoretical framework is still missing. Aiming to bridge this gap, in this paper, we prove generalization bounds for SGD under the assumption that its trajectories can be well-approximated by a *Feller process*, which defines a rich class of Markov processes that include several recent SDE representations (both Brownian or heavy-tailed) as its special case. We show that the generalization error can be controlled by the *Hausdorff dimension* of the trajectories, which is intimately linked to the tail behavior of the driving process. Our results imply that heavier-tailed processes should achieve better generalization; hence, the tail-index of the process can be used as a notion of "capacity metric". We support our theory with experiments on deep neural networks illustrating that the proposed capacity metric accurately estimates the generaliza-

Brownian motion

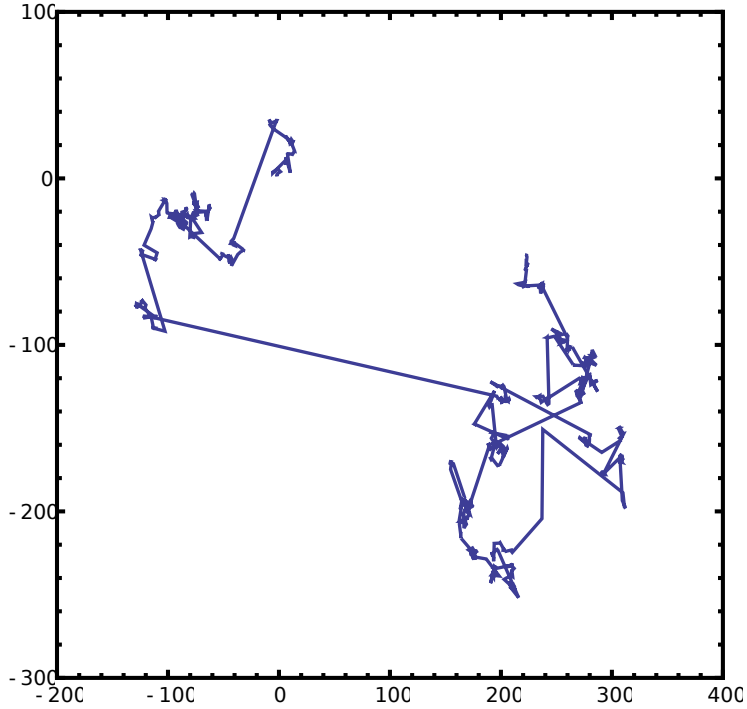
light-tailed



(exponential time to escape basins)

Levy flight

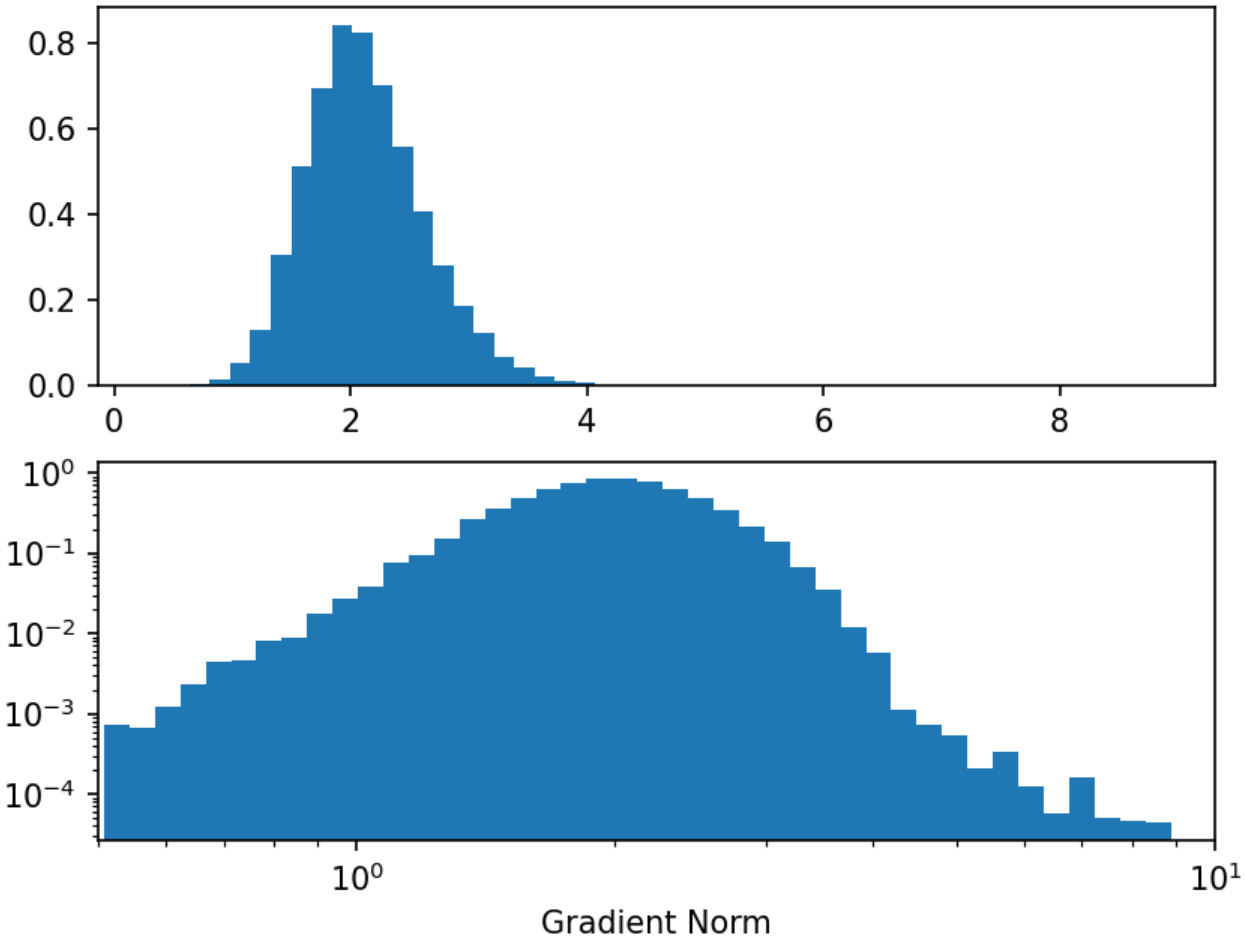
heavy-tailed



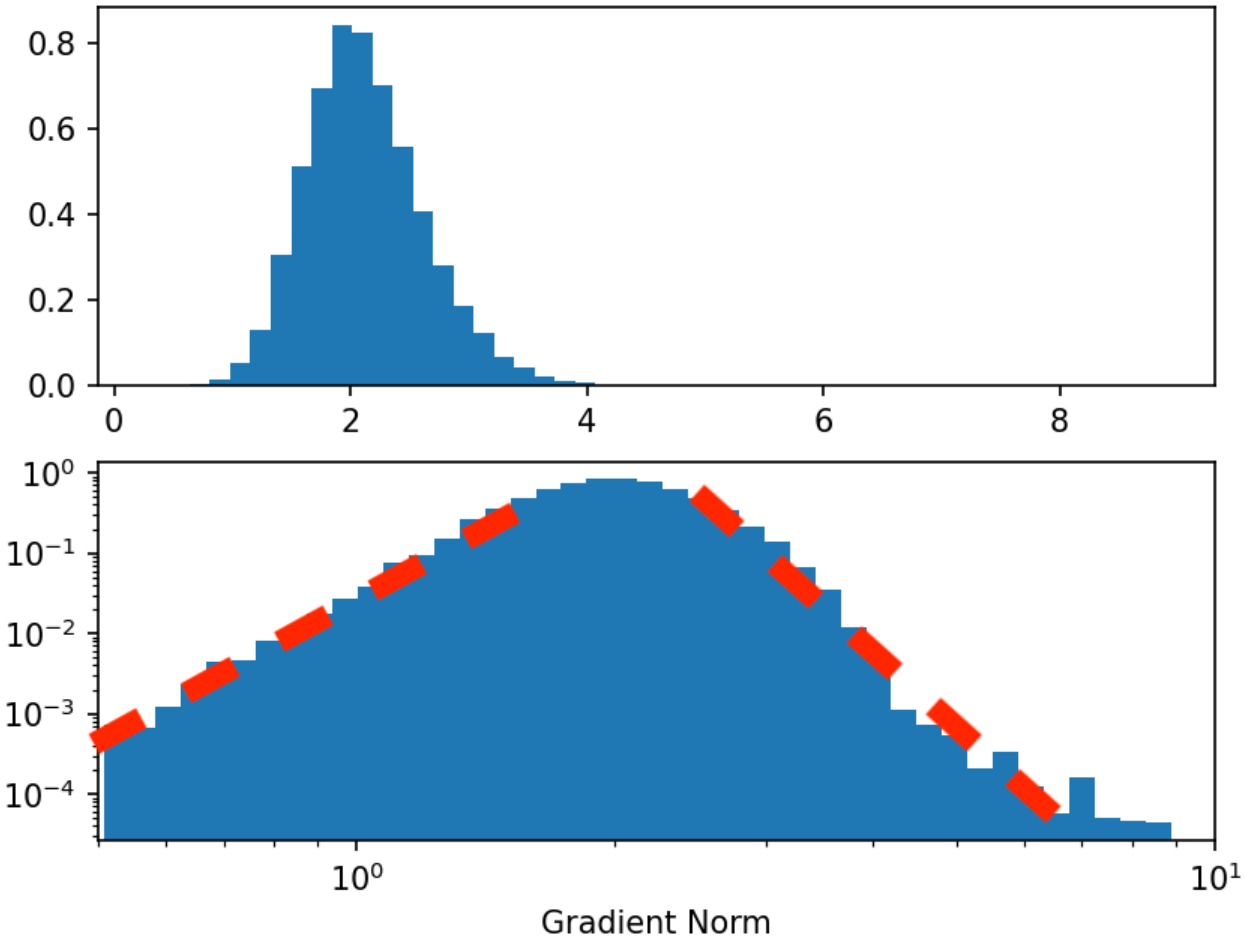
(polynomial time to escape basins; Nguyen et al., 2019)

The argument only works with continuous processes; what about actual SGD?

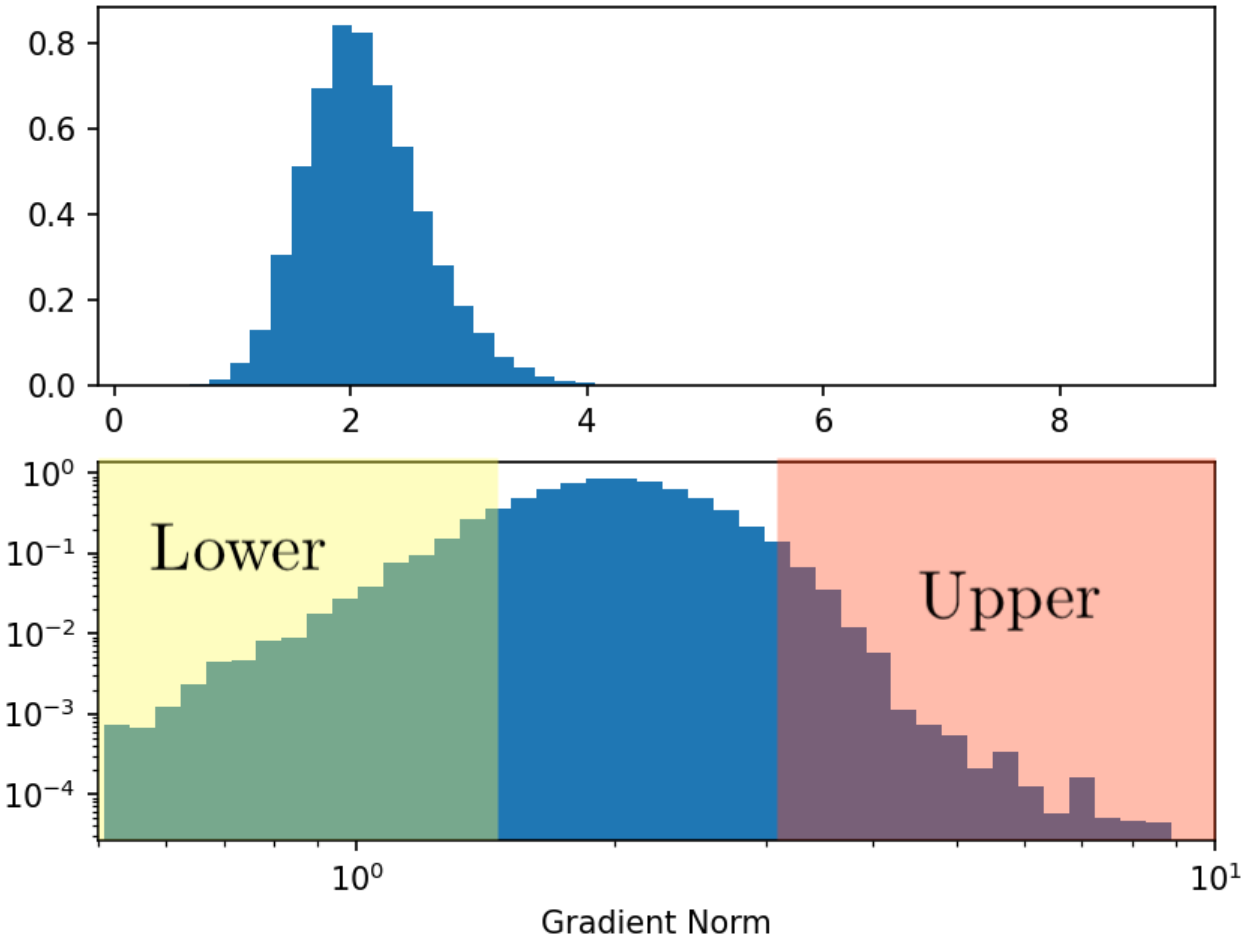
An Important Distinction



An Important Distinction



An Important Distinction



There are *two* kinds of power laws: **lower** and **upper**.

$$\mathbb{P}(X < x) \sim cx^{\alpha} \text{ or } \mathbb{P}(X > x) \sim cx^{-\beta}$$

Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers

Liam Hodgkinson [1] **Umut Şimşekli** [2] **Rajiv Khanna** [3] **Michael W. Mahoney** [1]

Abstract

Despite the ubiquitous use of stochastic optimization algorithms in machine learning, the precise impact of these algorithms and their dynamics on generalization performance in realistic non-convex settings is still poorly understood. While recent work has revealed connections between generalization and heavy-tailed behavior in stochastic optimization, this work mainly relied on continuous-time approximations; and a rigorous treatment for the original discrete-time iterations is yet to be performed. To bridge this gap, we present novel bounds linking generalization to the *lower tail exponent* of the transition kernel associated with the optimizer around a local minimum, in *both* discrete- and continuous-time settings. To achieve this, we first prove a data- and algorithm-dependent generalization bound in terms of the celebrated Fernique–Talagrand functional applied to the trajectory of the optimizer. Then, we specialize this result by exploiting the Markovian structure of stochastic optimizers, and derive bounds in terms of their (data-dependent) transition kernels. We support our theory with empirical results from a variety of neural networks, showing correlations between generalization error and lower tail exponents.

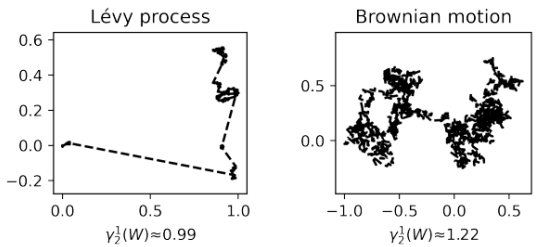


Figure 1: Discrete sample path approximations of a heavy-tailed α-stable Lévy process ($\alpha = 1.5$), and standard Brownian motion. Estimates of our normalized Fernique–Talagrand functional $\gamma_2^1(\cdot)$ is reported under each figure (see Section 2.3). Observe this functional is reduced with smaller tail index and "tighter clustering" of the trajectory.

1. Introduction

surprising generalization ability of stochastic gradient descent (SGD) and its various extensions for non-convex problems — most recently in the context of neural networks and deep learning. Classical convex optimization-centric approaches fail to explain this phenomenon.

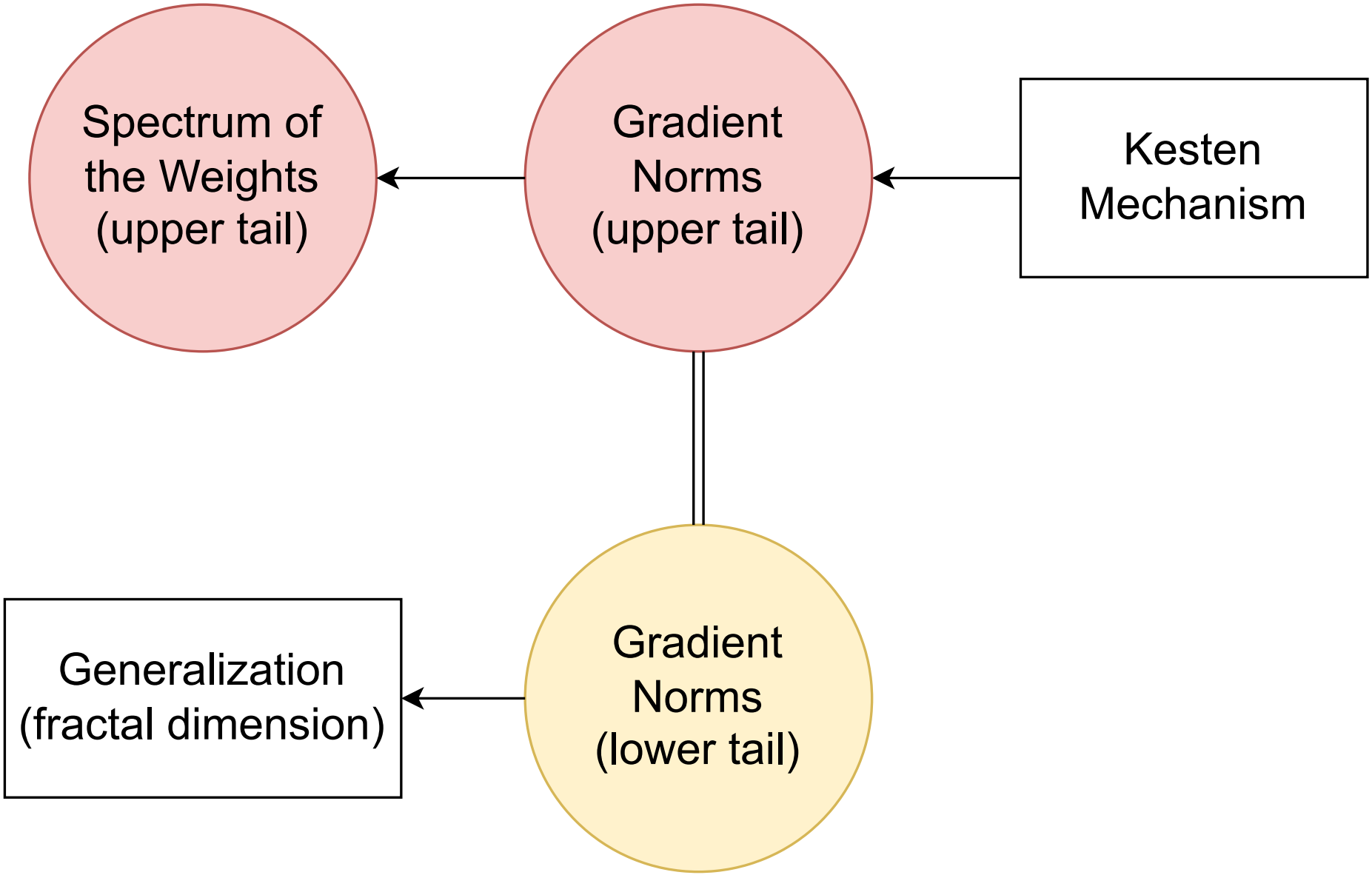
There has been an increasing number of attempts for developing generalization bounds for non-convex learning settings. This work has approached the problem from different perspectives, such as information theory, compression/sparsity/intrinsic dimension, or implicit (algorithmic) regularization (details to be provided in Section 1.2). Among these approaches, a promising direction has been to consider *optimization trajectories*, rather than single point estimates obtained during (or at the end of) the optimiza-

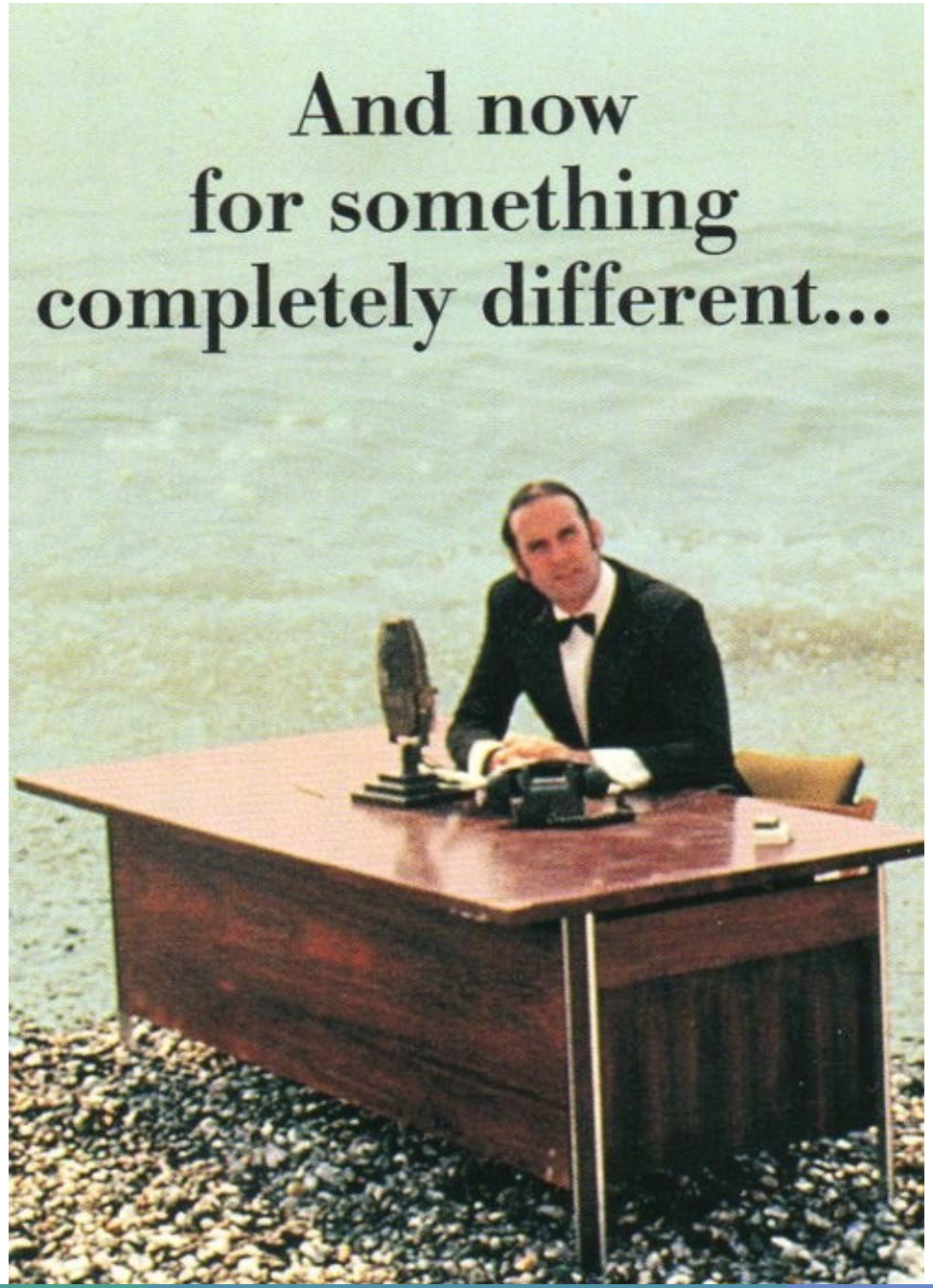
Theorem

If the optimization steps have a **lower power law** with exponent α in the neighbourhood of an optimum, then

$$\text{generalization gap} \leq c\alpha\sqrt{\frac{\log n}{n}}.$$

- Sachs, S., van Erven, T., Hodgkinson, L., Khanna, R., & Şimşekli, U. (2023). Generalization Guarantees via Algorithm-dependent Rademacher Complexity.
- Dupuis, B., Deligiannidis, G., & Şimşekli, U. (2023). Generalization bounds with data-dependent fractal dimensions.
- Lim, S. H., Wan, Y., & Simsekli, U. (2022). Chaotic regularization and heavy-tailed limits for deterministic gradient descent.


Spectrum of the Weights (upper tail)
Gradient Norms (upper tail)
Kesten Mechanism
Generalization (fractal dimension)
Gradient Norms (lower tail)


And now
for something
completely different...

Overparameterization

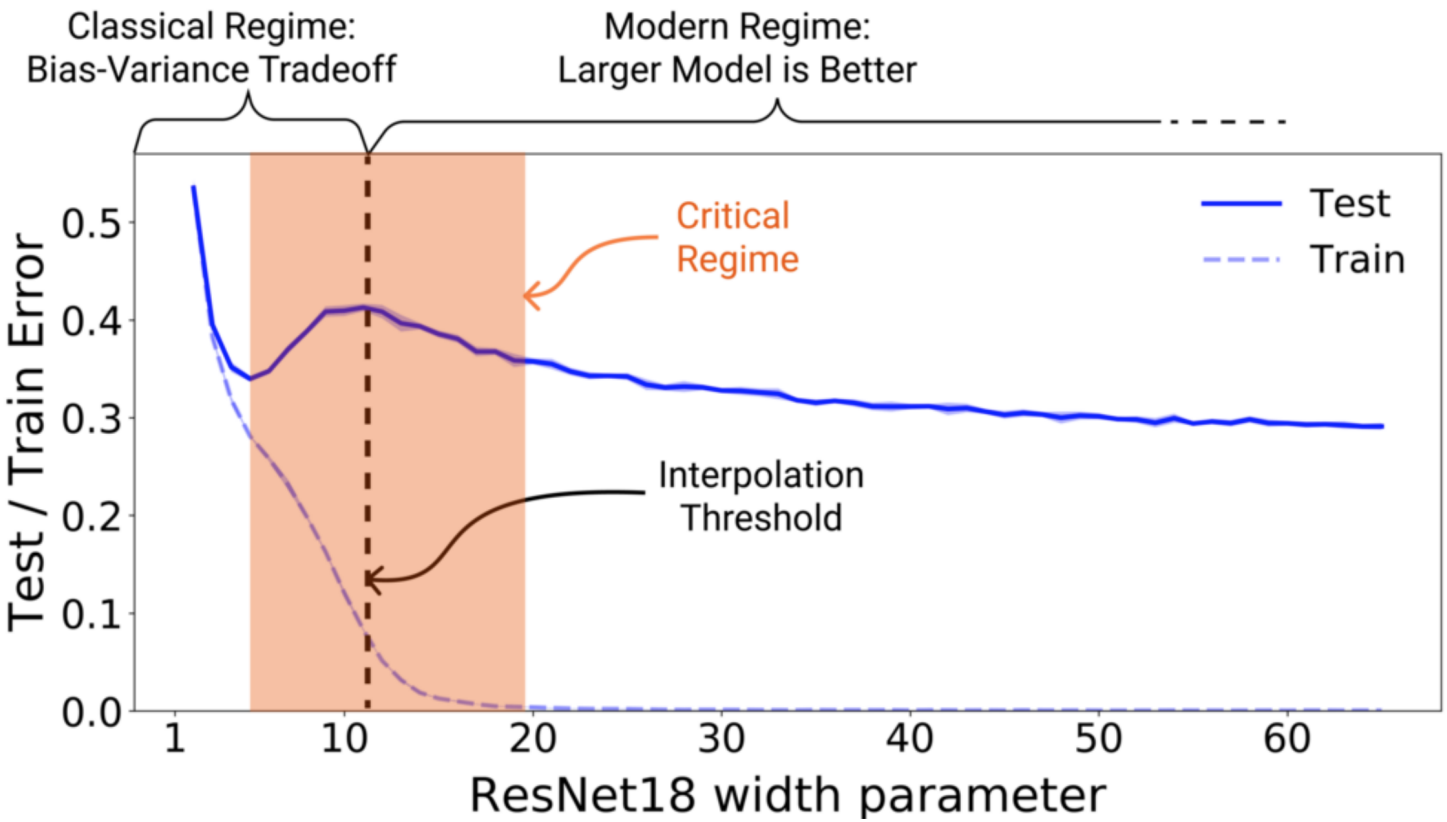
Overparameterized Models

The most performant models are *often* overparameterized.

Model	Dataset	($\log_{10}$ scale) N	d	Test Accuracy
`ResNet18`	`CIFAR-10`	5.7	7.0	93%
`WRN-28-10`	`SVHN`	6.8	7.6	98%
`ViT-E`	`ImageNet-1k`	9.1	9.6	91%
`EFL`	`SNLI`	6.2	8.6	93%
`ResNet34`	`Chaoyang` (noisy)	4.4	7.8	83%
`FiLM`	`ETT` (noisy)	4.0	6.0	—

...and have virtually zero error on training set.

Double Descent



Nakkiran, Preetum, et al. *Deep double descent: Where bigger models and more data hurt*. Journal of Statistical Mechanics: Theory and Experiment 2021.12 (2021): 124003.

Monotonicity and Double Descent in Uncertainty Estimation with Gaussian Processes

Liam Hodgkinson [1] **Chris van der Heide** [2] **Fred Roosta** [3 4 5] **Michael W. Mahoney** [5 6 7]

Abstract

Despite their importance for assessing reliability of predictions, uncertainty quantification (UQ) measures for machine learning models have only recently begun to be rigorously characterized. One prominent issue is the *curse of dimensionality*: it is commonly believed that the marginal likelihood should be reminiscent of cross-validation metrics and that both should deteriorate with larger input dimensions. We prove that by tuning hyperparameters to maximize marginal likelihood (the empirical Bayes procedure), the performance, as measured by the marginal likelihood, *improves monotonically* with the input dimension. On the other hand, we prove that cross-validation metrics exhibit qualitatively different behavior that is characteristic of *double descent*. Cold posteriors, which have recently attracted interest due to their improved performance in certain settings, appear to exacerbate these phenomena. We verify

tion of error and other methods for inverse uncertainty problems typically apply Monte Carlo methods under a Bayesian framework (Zhang, 2021). However, the large-scale nature of many problems of interest results in significant computational challenges. One of the most successful approaches for solving inverse uncertainty problems is the use of *Gaussian processes* (GP) (Rasmussen & Williams, 2006). This is now frequently used for many predictive tasks, including time-series analysis (Roberts et al., 2013), regression and classification (Rasmussen & Williams, 2006; Williams & Barber, 1998). GPs are also valuable in deep learning theory due to their appearance in the infinite-width limits of Bayesian neural networks (Jacot et al., 2018; Neal, 1996).

A prominent feature of modern ML tasks is their large number of attributes: for example, in computer vision and natural language tasks, input dimensions can easily scale into the tens of thousands. This is concerning in light of the prevailing theory that GP performance often deteriorates in higher input dimensions. This *curse of dimensionality* for GPs has been rigorously demonstrated through error estimates for

The Interpolating Information Criterion for Overparameterized Models

Liam Hodgkinson[1], **Chris van der Heide**[2], **Robert Salomone**[3], **Fred Roosta**[4] **and Michael W. Mahoney**[5]

[1]*School of Mathematics and Statistics, University of Melbourne.*
e-mail: lhodgkinson@unimelb.edu.au

[2]*Department of Electrical and Electronic Engineering, University of Melbourne.*
e-mail: chris.vdh@gmail.com

[3]*Centre for Data Science, Queensland University of Technology.*
e-mail: robert.salomone@qut.edu.au

[4]*CIRES and School of Mathematics and Physics, University of Queensland.*
e-mail: fred.roosta@uq.edu.au

[5]*ICSI, LBNL, and Department of Statistics, University of California, Berkeley.*
e-mail: mahoney@stat.berkeley.edu

Abstract: The problem of model selection is considered for the setting of interpolating estimators, where the number of model parameters exceeds the size of the dataset. Classical information criteria typically consider the large-data limit, penalizing model size. However, these criteria are not appropriate in modern settings where overparameterized models tend to perform well. For any overparameterized model, we show that there exists a dual underparameterized model that possesses the same marginal likelihood, thus establishing a form of *Bayesian duality*. This enables more classical methods to be used in the overparameterized setting, revealing the *Interpolating Information Criterion*, a measure of model quality that naturally incorporates the choice of prior into the model selection. Our new information criterion accounts for prior misspecification, geometric and spectral properties of the model, and is numerically consistent with known empirical and theoretical behavior in this regime.

Interpolating Information Criterion

Theorem

Under mild conditions, if the loss is σ^2-subgaussian, the **expected test error** in a neighbourhood of the interpolating solution is bounded above by

$$\frac{c}{2}\mathbf{IIC} + \sigma^2 + n^{-1}\log(\delta^{-1}) + \text{const.} + \mathcal{O}(n^{-2}),$$

with probability at least $1 - \delta$, where IIC is our *Interpolating Information Criterion*.

Interpolating Information Criterion

- Hodgkinson, L., Van Der Heide, C., Roosta, F., & Mahoney, M. W. (2023). **Monotonicity and double descent in uncertainty estimation with Gaussian processes**.
- Hodgkinson, L., van der Heide, C., Salomone, R., Roosta, F., & Mahoney, M. W. (2023). **A PAC-Bayesian Perspective on the Interpolating Information Criterion.**

$$\mathrm{IIC} = \log R(\theta^\star) + \text{curvature} + \frac{1}{n}\textbf{log det(Gram matrix of NTK)}.$$

Neural Tangent Kernel

The neural tangent kernel is given by

$$k_{\mathrm{NTK}}(x, y) = \nabla_\theta f(x, \theta) \cdot \nabla_\theta f(y, \theta).$$

The NTK matrix as it appears in the IIC is its corresponding *Gram matrix* over the training set

$$\begin{aligned} \text{Gram matrix of NTK} &= (k_{\mathrm{NTK}}(x_i, x_j))_{i,j=1}^{n} \\ &= JJ^\top \in \mathbb{R}^{N \times N}. \end{aligned}$$

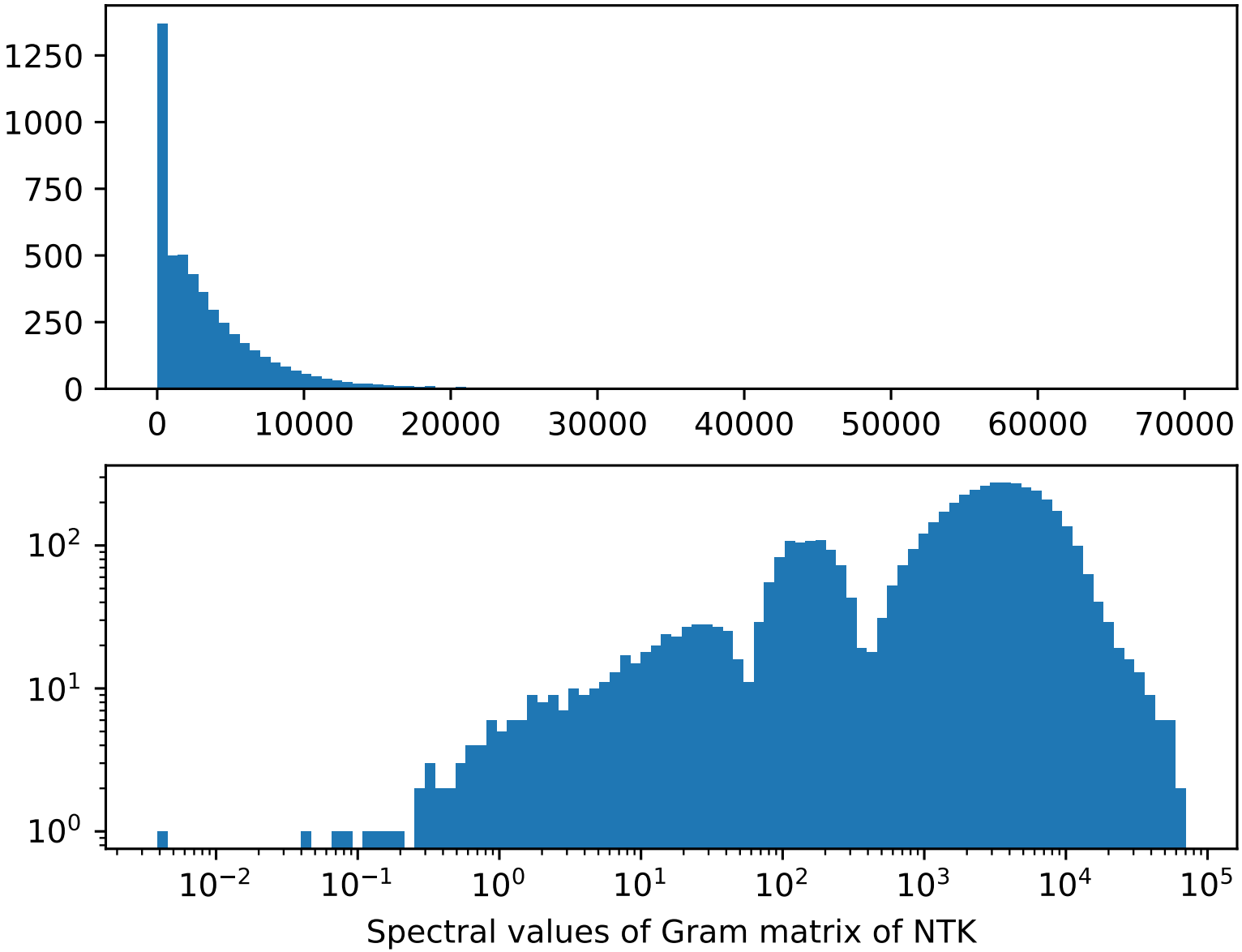
The Spectrum Appears!

$$\text{IIC} = \text{log-regularizer} + \text{curvature} + \frac{1}{n}\textbf{log det(Gram matrix of NTK)}.$$

Lemma

If spectrum of A has a lower tail exponent α, then

$$\log\det A \propto \alpha.$$


1250
1000
750
500
250
0
0
10000
20000
30000
40000
50000
60000
70000
10^2
10^1
10^0
10^{-2}
10^{-1}
10^0
10^1
10^2
10^3
10^4
10^5
Spectral values of Gram matrix of NTK

The Origin

When **overparameterized**, the **maximum entropy distribution** for the NTK eigenspectrum that **minimizes the IIC** is

$$p(\lambda) \sim c\lambda^{\alpha}.$$

Xie, Z., Tang, Q. Y., Cai, Y., Sun, M., & Li, P. (2022). On the Power-Law Hessian Spectrums in Deep Learning.

The Gram matrix of the NTK has a lower power law.

Can we get everything else from here?

- **Scaling laws:** As the lower tail of the Gram matrix becomes heavier, training takes *longer*.

  Velikanov, M., & Yarotsky, D. (2021). Explicit loss asymptotics in the gradient descent training of neural networks. Advances in Neural Information Processing Systems, 34, 2570-2582.

- **Dynamics:** If the Gram matrix of the NTK has lower power law, then so does the gradient norm.

Spectrum of Weights

The spectrum of the weights captures much of the same information.

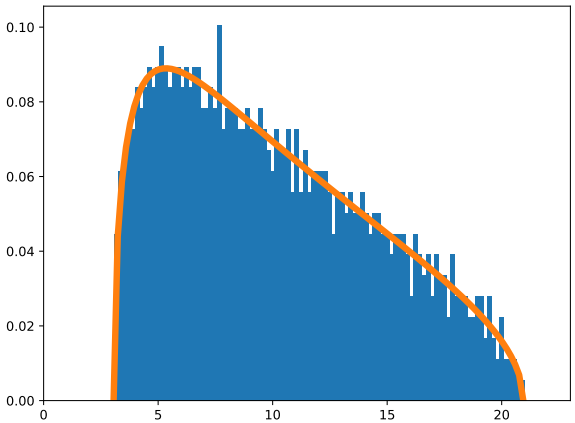
- If Gram matrix of the NTK has a lower power law, it is **near-singular**.
- A **near-singular** Gram matrix implies a near-singular **optimization problem**.
- By implicit function theorem,

$$\Delta W \approx J^{\dagger} X,$$

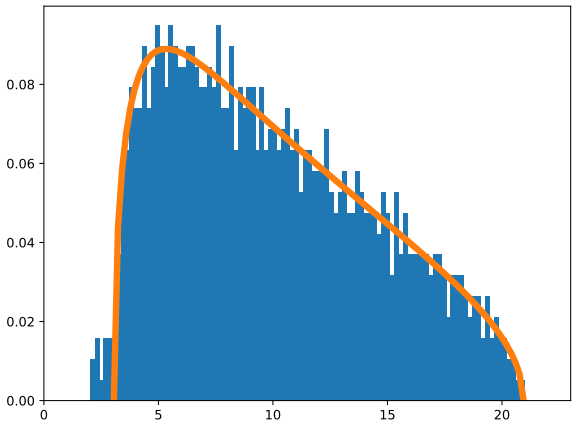
so weights accumulate an upper power law at the end of training.

Five phase taxonomy

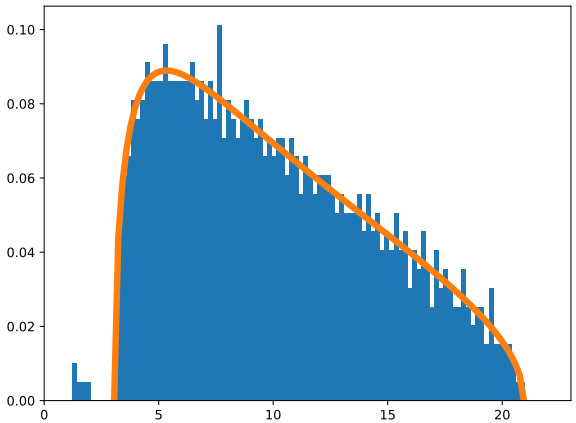
Random-like



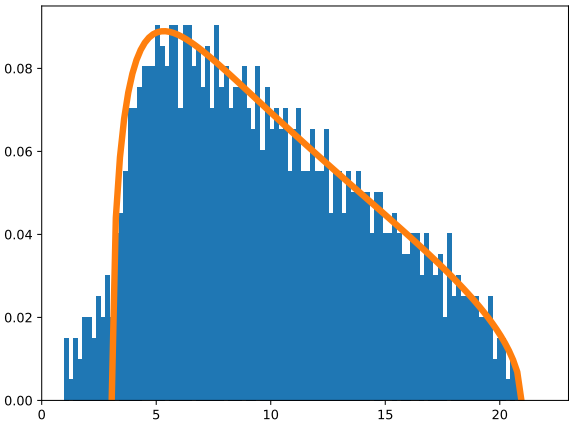
Bleeding out



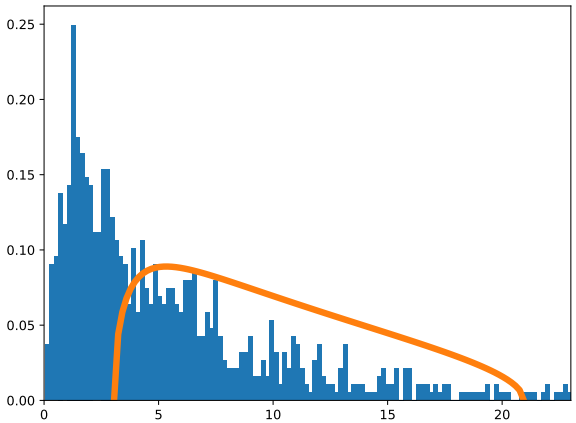
Bulk + spike

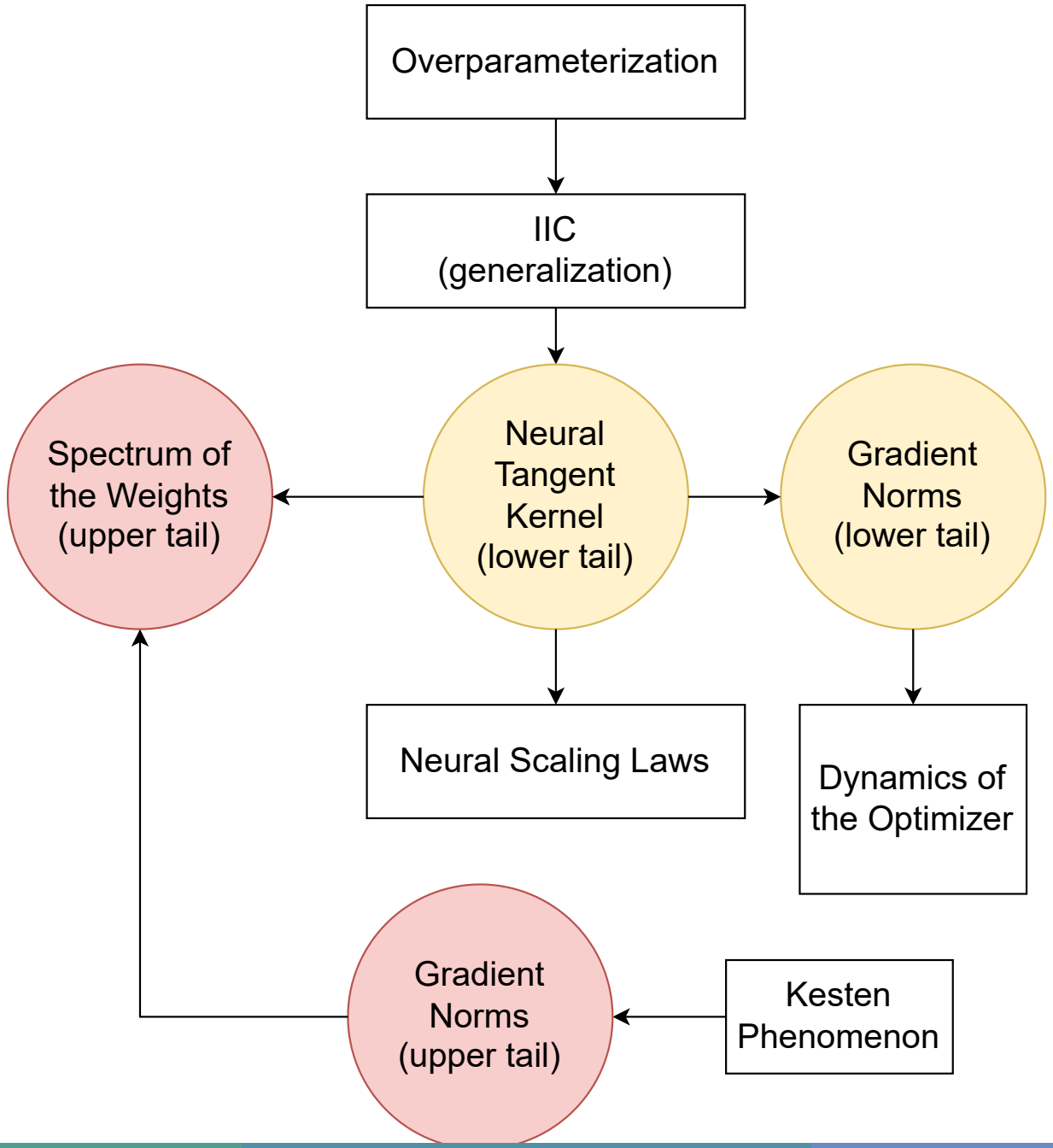


Bulk + decay



Power law




Overparameterization
IIC
(generalization)
Spectrum of
the Weights
(upper tail)
Neural
Tangent
Kernel
(lower tail)
Gradient
Norms
(lower tail)
Neural Scaling Laws
Dynamics of
the Optimizer
Gradient
Norms
(upper tail)
Kesten
Phenomenon

The Future

Everything is connected.

Two powerful tools for examining heavy tails in ML:

- Interpolating Information Criterion
- Fractal dimension bounds

Gram matrix of NTK is a key object, but there may be something better...

Exponentially-truncated power laws?