Heavy-Tailed Universality Predicts Trends in Test Accuracies for Very Large Pre-Trained Deep Neural Networks

Anonymous Authors[1]

Abstract

Given two or more Deep Neural Networks (DNNs) with the same or similar architectures, and trained on the same dataset, but trained with different solvers, parameters, hyper-parameters, regularization, etc., can we predict which DNN will have the best test accuracy, and can we do so without peeking at the test data? In this paper, we show how to use a new Theory of Heavy-Tailed Self-Regularization (HT-SR) to answer this. HT-SR suggests, among other things, that modern DNNs exhibit what we call Heavy-Tailed Mechanistic Universality (HT-MU), meaning that the correlations in the layer weight matrices can be fit to a power law with exponents that lie in common Universality classes from Heavy-Tailed Random Matrix Theory (HT-RMT). From this, we develop a Universal capacity control metric that is a weighted average of these PL exponents. Rather than considering small toy NNs, we examine over 50 different, large-scale pre-trained DNNs, ranging over 15 different architectures, trained on ImagetNet, each of which has been reported to have different test accuracies. We show that this new capacity metric correlates very well with the reported test accuracies of these DNNs, looking across each architecture (VGG16/.../VGG19, ResNet10/.../ResNet152, etc.). We also show how to approximate the metric by the more familiar Product Norm capacity measure, as the average of the log Frobenius norm of the layer weight matrices. Our approach requires no changes to the underlying DNN or its loss function, it does not require us to train a model (although it could be used to monitor training), and it does not even require access to the ImageNet data.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

We are interested in the following general question.

• Given two or more Deep Neural Networks (DNNs) with the same or similar architectures, and trained on the same dataset, but trained with different solvers, parameters, hyper-parameters, regularization, etc., can we predict which DNN will have the best test accuracy, and can we do so without peeking at the test data?

This question is both theoretical and practical. Theoretically, solving this would help to understand why this class of machine learning (ML) models performs as well as it does in certain classes of applications. Practically, there are many motivating examples. Here are two.

• **Automating architecture search.** Developing DNN models often requires significant architecture engineering, so there is interest in automating the design of DNNs. Current methods can produce a series of DNNs subject to given general architecture constraints, but the models must be evaluated using cross validation (CV). DNNs have so many adjustable parameters that even when using CV it is possible to leak information from the test sets into the training data, thus producing brittle, non-robust models. It is thus of interest to have design principles and quality metrics that do not depend on the test data and/or the labels.

• **Fine-Tuning Pre-trained Models.** Since one often does not have enough labeled data to train a large DNN from scratch, many modern engineering solutions can re-use widely-available pre-trained DNNs, fine-tuning them on smaller data sets. This technique often works extremely well for visual tasks, using DNNs pre-trained on ImageNet; and recently it has become feasible for complex natural language processing (NLP) tasks. Sometimes, however, these fine-tuned models become brittle and non-robust—due to overtraining, because information leaks from the test set into the training data. Here, it would also be very helpful to be able to fine-tune large, pre-trained DNNs without needing to peek at the test data.

To predict trends in the generalization accuracy of a series of DNN architectures, VC-like theories offer theoretical bounds on the generalization accuracy. Practically, such capacity metrics can guide the theoretical development of

new regularizers for traditional ML optimization problems (e.g., counterfactual expected risk minimization (Swaminathan & Joachims, 2015)), but the bounds themselves are far too loose to be used directly. Moreover, since the early days of NN research, it was known that VC theory could (probably) not be directly applied to the seemingly widely non-convex optimization problem implicitly posed by NNs. (This has caused some researchers to suggest we need to rethink regularization in DNNs entirely.)

In light of this, Liao et al. (Liao et al., 2018) used an appropriately-scaled, data-dependent Product Norm capacity control metric to bound the worst-case generalization error for several small (non production-quality, but still interesting) DNN models, and they showed that the bounds are remarkably tight. There is, in fact, a large body of work on norm-based capacity control metrics, both recent, e.g., (Liao et al., 2018; Soudry et al., 2017; Poggio et al., 2018) and (Neyshabur et al., 2014; 2015; 2017a; Bartlett et al., 2017; Yoshida & Miyato, 2017; Kawaguchi et al., 2017; Neyshabur et al., 2017b; Arora et al., 2018b;a; Zhou & Feng, 2018), as well as much older (Bartlett, 1997; Mahoney & Narayanan, 2009). Much of this work has been motivated by the observation that parameter counting and more traditional VC-based bounds tend to lead to vacuous results for modern state-of-the-art DNNs, e.g., since modern DNNs are heavily over-parameterized and depend so strongly on the training data.

As with most theoretical studies, Liao et al.'s approach and intent differ greatly from ours. They seek *worst-case* complexity bounds, motivated to reconcile discrepancies with more traditional statistical learning theory, and they apply it to quite small NNs. To address our main question, we seek an *average-case* or *typical case* (for realistic problems) complexity metric, viable in production to guide the development of better DNNs at scale. Bounding a toy model does not necessarily mean that the individual weight matrix norms in production-quality DNNs will be directly comparable. In particular, it does not mean that we can directly compare the individual weight matrix norms across layers in different, and more complex, architectures. Also, Liao et al. had to modify the DNN optimization loss function. This means that their approach can not be tested/evaluated on any existing pre-trained DNN architecture, e.g., the VGG and ResNet models, widely-used today in industry. Still, their results do suggest that a Product Norm may work well as a *practical* capacity metric for large, and perhaps even pre-trained, production-quality DNNs. To predict trends in the test accuracies, one needs some more *Universal* empirical metric that transfers across DNN architectures.

Recent work by Martin and Mahoney (Martin & Mahoney, 2018a) provides a Universal empirical metric that characterizes the amount of *Implicit Self-Regularization* and, accordingly, the generalization capacity, for a wide range of pre-trained DNNs. The metric involves the power law (PL) exponents, α, of individual layer weight matrices, $\mathbf{W}$, as determined by fitting the Empirical Spectral Density (ESD), $\rho(\lambda)$, to a PL distribution. Looking in detail at a series of models, like AlexNet, VGG, ResNet, etc, they observe that the (linear) layer weight matrices almost always follow a PL distribution, and the fitted PL exponents nearly all lie within a universal range $\alpha \in [2, 5]$. Further, analysis of a small model (MinAlexNet) demonstrates that smaller PL exponents α correspond to better generalization. Subsequent work (Martin & Mahoney, 2018b) demonstrated HT behavior in nearly every pre-trained architecture studied, e.g., across nearly 7500 layer weight matrices (and 2D feature maps), including DNNs pre-trained for computer vision tasks on ImageNet, and for several different NLP tasks.

When one observes good empirical PL fits of the ESDs of the correlations of layer weight matrices, we say the DNN *exhibits Heavy-Tailed (HT) behavior*. Motivated by these empirical observations, and using the Universality properties of Heavy-Tailed Random Matrix Theory (HT-RMT), Martin and Mahoney developed a theory of Heavy-Tailed Self-Regularization (HT-SR) for DNNs (Martin & Mahoney, 2017; 2018a). We build on and extend that theory here.

In Statistical Physics, Universality of PL exponents is very special, and it suggests the presence of a deeper, underlying, *Universal mechanism* driving the system dynamics (Sornette, 2006; Bouchaud & Potters, 2003). It is this *Heavy Tailed Mechanistic Universality* (HT-MU), as well call it, that originally motivated our study. HT-MU applies to the analysis of complicated systems, including many physical systems, traditional NNs (Engel & den Broeck, 2001; Nishimori, 2001), and even models of the dynamics of actual spiking neurons. Indeed, the dynamics of learning in DNNs seems to resemble a system near a phase transition, such as the phase boundary of spin glass, or a system displaying Self Organized Criticality (SOC), or a Jamming transition (Geiger et al., 2018; Spigler et al., 2018). Of course, we can not say which mechanism, if any, is at play. Instead, we use the machinery of HT-RMT as a stand-in for a generative model of the weight matrices in DNNs, to catalog and model the HT behavior of DNNs.[1] This Universality *suggests* that we look for a *Universal Capacity Control Metric*[2] to address

[1] Perhaps the most well-known Universality in RMT is associated with the Gaussian Universality class, where the sum of many random variables drawn from a wide range of distributions is "approximately Gaussian," e.g., in the sense that the sum approaches a suitably-normalized Gaussian distribution. As briefly reviewed in Section 2, HT Universality makes analogous (but, admittedly, more complicated) statements for random variables drawn from distributions in which the tails decay more slowly than those in the Gaussian Universality class (Martin & Mahoney, 2018a).

[2] To be clear, this metric is Universal, not in the sense that it will apply "universally" to every possible DNN, but in the Statistical

our main question.

Our main results are the following.

- We introduce a new methodology to analyze the performance of large-scale pre-trained DNNs, using a phenomena observed in HT-SR Theory that we call Heavy-Tailed Mechanistic Universality (HT-MU). We construct a Universal capacity control metric to predict average DNN test performance. This metric is a weighted average of layer PL exponents, $\hat{\alpha}$, weighted by the log[3] of the Spectral norm (i.e., maximum eigenvalue λ^{max}) of layer correlation matrices:

$$\hat{\alpha} = \sum_{l \in L} \alpha_l \log \lambda_l^{max}.$$

- We apply our Universal capacity control metric $\hat{\alpha}$ to a wide range of large-scale pre-trained production-level DNNs, including the VGG and ResNet series of models, as well as many others. This Universal metric correlates very well with the reported average test accuracies across many series of pre-trained DNNs.
- We derive a relation between our Universal capacity control metric $\hat{\alpha}$ and the well known Product Norm capacity control metric, i.e, in the form of the average log of the squared Frobenius norm:

$$\langle \log \|\mathbf{W}\|_F^2 \rangle \approx \alpha \log \lambda^{max}.$$

We also show that such norm-based capacity control metrics also correlate well with the average test accuracy in large-scale production-level pre-trained DNNs.

For both our Universal metric and the Product Norm metric, our empirical results are, to our knowledge, the first time such theoretical capacity metrics have been reported to predict (trends in) the test accuracy for *pre-trained production-level* DNNs. In particular, this illustrates the usefulness of these norm-based metrics beyond smaller models such as MNIST, CIFAR10, and CIFAR100. Our results, including for both our Universal metric and the Product Norm metric we consider, can be reproduced with the `WeightWatcher` package[4]; and our results suggest that our "practical theory" approach is fruitful more generally for engineering good algorithms for realistic large-scale DNNs.

2. Overview of Heavy-Tailed Self-Regularization

We review Martin and Mahoney's Theory of Heavy-Tailed Self-Regularization (HT-SR) (Martin & Mahoney, 2018a).

Write the Energy Landscape (or optimization function) for a typical DNN with L layers, with activation functions $h_l(\cdot)$, and with $N \times M$ weight matrices $\mathbf{W}_l$ and biases $\mathbf{b}_l$, as:

$$E_{DNN} = h_L(\mathbf{W}_L \cdot h_{L-1}(\mathbf{W}_{L-1} \cdot h_{L-2}(\cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L).$$

Typically, this model would be trained on some labeled data $\{d_i, y_i\} \in \mathcal{D}$, using Backprop, by minimizing the loss $\mathcal{L}$. For simplicity, we do not indicate the structural details of the layers (e.g., Dense or not, Convolutions or not, Residual/Skip Connections, etc.). Each layer is defined by one or more layer 2D weight matrices $\mathbf{W}_l$, and/or the 2D feature maps $\mathbf{W}_{l,i}$ extracted from 2D Convolutional (Conv2D) layers. (We have not yet analyzed LSTM or other complicated Layers.) A typical modern DNN may have anywhere between 5 and 5000 2D layer matrices.[5]

Heavy-Tailed Empirical Spectral Distributions. In the HT-SR Theory, we analyze the eigenvalue spectrum (the ESD) of the associated correlation matrices (Martin & Mahoney, 2018a). From this, we can characterize the amount and form of correlation, and therefore implicit self-regularizartion, present in the DNN's weight matrices. For each layer weight matrix, construct the associated $M \times M$ (uncentered) correlation matrix $\mathbf{X}$. Dropping the L and l, i indices, we have

$$\mathbf{X} = \frac{1}{N} \mathbf{W}^T \mathbf{W}.$$

If we compute the eigenvalue spectrum of $\mathbf{X}$, i.e., λ_i such that $\mathbf{X}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, then the ESD of eigenvalues, $\rho(\lambda)$, is just a histogram of the eigenvalues, formally written as

$$\rho(\lambda) = \sum_{i=1}^{M} \delta(\lambda - \lambda_i). \tag{1}$$

Using HT-SR Theory, we can characterize the *correlations* in a weight matrix by examining its ESD, $\rho(\lambda)$. It can be well-fit to a power law (PL) distribution, given as

$$\rho(\lambda) \sim \lambda^{-\alpha}, \tag{2}$$

which is (at least) valid within a bounded range of eigenvalues $\lambda \in [\lambda^{min}, \lambda^{max}]$. We can determine α by fitting the ESD to a PL, using the commonly accepted Maximum Likelihood (MLE) method of Clauset et al. (Clauset et al., 2009; Alstott et al., 2014). This method works very well for exponents between $\alpha \in (2, 4)$, and it is adequate, although imprecise, for smaller and larger α (Newman, 2005).

Physics sense (Sornette, 2006; Bouchaud & Potters, 2003) that it should apply to matrices within/across HT "Universality" classes.

[3] Throughout, we use log base 10.

[4] https://pypi.org/project/WeightWatcher/

[5] For each Linear Layer, we get a single $(N \times M)$ (real-valued) 2D weight matrix, denoted $\mathbf{W}_l$, for layer l. This includes Dense or Fully-Connected (FC) layers, as well as 1D Convolutional (Conv1D) layers, Attention matrices, etc. We ignore the bias here terms $\mathbf{b}_l$ in this analysis. Let the aspect ratio be $Q = \frac{N}{M}$, with $Q \geq 1$. For the Conv2D layers, we have a 4-index Tensor, of the form $(N \times M \times c \times d)$, consisting of $c \times d$ 2D feature maps of shape $(N \times M)$. We extract $n_l = c \times d$ 2D weight matrices $\mathbf{W}_{l,i}$, one for each feature map $i = [1, \ldots, n_l]$ for layer l.

The original work on HT-SR Theory (Martin & Mahoney, 2018a) considered NNs including AlexNet and InceptionV3 (as well as DenseNet, ResNet, and VGG), and it showed that for nearly every $\mathbf{W}$, the (bulk and tail) of the ESDs can be fit to a PL and the PL exponents α nearly all lie within the range $\alpha \in (1.5, 5)$. Moreover, smaller exponents α are correlated with more implicit self-regularization and, correspondingly, better generalization (Martin & Mahoney, 2018a). Subsequent work (Martin & Mahoney, 2018b) has shown that these results are ubiquitous. For example, upon examining nearly 10,000 layer weight matrices $\mathbf{W}_{l,i}$ across over 50 different modern pre-trained DNN architectures, the ESD of nearly every $\mathbf{W}$ layer matrix can be fit to a PL: $70-80\%$ of the time, the fitted PL exponent α lies in the range $\alpha \in (2, 4)$; and $10-20\%$ of the time, the fitted PL exponent α lies in the range $\alpha < 2$. For example, see Figure 1 for a histogram of results for ca. 7500 weight matrices from ImageNet. Of course, there are exceptions: in any real DNN, the fitted α may range anywhere from ~ 1.5 to 10 or higher (Martin & Mahoney, 2018b) (and, of course, larger values of α may indicate that the PL is not a good model for the data). Still, overall, in nearly all large, pre-trained DNNs, the correlations in the weight matrices exhbit a remarkable Universality, being both Heavy Tailed, and having small—but not too small—PL exponents.

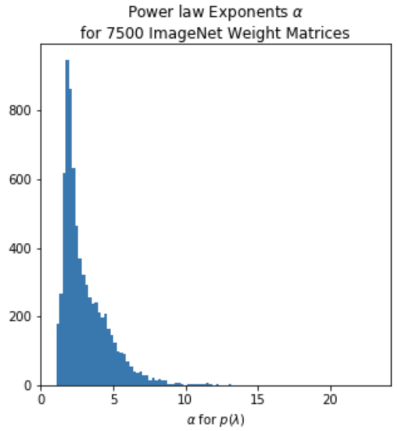


Figure 1. Histogram of PL exponents, α, fit on ca. 7500 Linear and Conv2D Layers from ImageNet. The vast majority have $\alpha \in (1.5, 5)$. Some Conv2D layers have smaller values of α; and larger values of α (up to ca. 20) exist but correspond to less reliable fits.

Heavy-Tailed Mechanistic Universality. Here, we consider the question: what does it *mean* to say that DNN weight matrices exhibit Universality? This answer to this—and its implications—depends on your perspective, in particular, as a Mathematician or a Physicist.

In Statistics and Applied Mathematics, Universality typically refers to properties of systems that can be modeled by random matrices. The justification is that certain system properties can be deduced, without requiring knowledge of system details, from a few global quantities that are then used as parameters to define a random matrix ensemble (Edelman & Rao, 2005; Edelman & Wang, 2013). Of course, the DNN weight matrices are not random matrices—they are strongly- correlated objects—so it may seem odd that we can apply RMT to characterize them.

In Statistical Physics, Universality refers to a different, but related, phenomena. It arises in systems with very strong correlations, at or near a critical point or phase transition. It is characterized by measuring experimentally certain "observables" that display HT behavior, with common—or Universal—PL exponents. More importantly, it indicates that a specific Universal mechanism drives the underlying physical process, e.g., Self Organized Criticality, directed percolation, etc. (Sornette, 2006; Bouchaud & Potters, 2003). For this reason, we refer to the Universality observed in HT-SR, i.e., in the ESDs of (pre-trtained) DNN weight matrices, as *Heavy-Tailed Mechanistic Universality (HT-MU)*.

For an illustration of what we mean by HT-MU, see Figure 2. When we observe HT behavior in $\mathbf{W}$, or rather its correlation matrix $\mathbf{X}$, we use HT-RMT as a generative model. We say that we *model* $\mathbf{W}$ *as if* it is a random matrix, $\mathbf{W}^{rand}(\mu)$, drawn from a Universality class of HT-RMT (i.e., VHT, MHT, or WHT, as defined below). Of course, we do not mean that $\mathbf{W}$ is itself random in any way. We simply use RMT as a stand-in generative model because the correlations in $\mathbf{W}^{rand}(\mu)$ resembles the correlations in $\mathbf{W}$. Indeed, RMT itself does not describe the behavior of a random matrix $\mathbf{W}$ per-se, but it characterizes its correlations (i.e., the eigenvalues of $\mathbf{X}$). Specifically, RMT describes the ESD, $\rho_{emp}(\lambda)$, its limiting, deterministic form $\rho_{\infty}(\lambda)$, as well as the finite-size scaling and fluctuations of the maximum eigenvalue, λ^{max} (Martin & Mahoney, 2018a). So, even though $\mathbf{W}$ is not random, we expect its ESD and maximum eigenvalue to behave *as if* they were drawn from some HT-random matrix $\mathbf{W}^{rand}(\mu)$.

Heavy-Tailed Random Matrix Theory. To characterize this HT-MU behavior, we use a HT variant of RMT and use HT random matrices to elucidate different Universality classes. Let $\mathbf{W}(\mu)$ be an $N \times M$ random matrix with entries chosen i.i.d. from

$$\mathbf{Pr}\left[W_{i,j}\right] \sim \frac{W_0^{\mu}}{|W_{i,j}|^{1+\mu}}, \tag{3}$$

where W_0 is the typical order of magnitude of $W_{i,j}$, and where $\mu > 0$. These HT matrix models were first introduced in the Statistical Physics literature, where they are called Lévy Matrices when $0 < \mu < 2$ (Cizeau & Bouchaud, 1994); see also (Bouchaud & Mézard, 1997; Burda et al.,

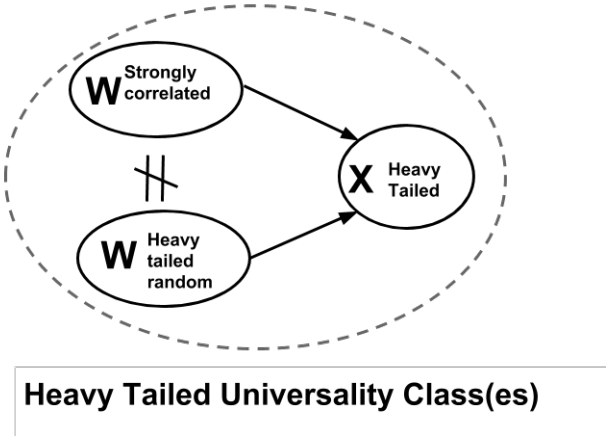


Figure 2. Illustration of our use of HT Universality. A random matrix $\mathbf{W}(\mu)$ with elements drawn i.i.d. from the HT distribution of Eqn. (3) and weight matrix $\mathbf{W}^{str.corr.}$ from a modern well-trained DNN that exhibits strong correlations each arise from different generative mechanisms, and they are different. Both exhibit similar Universality properties, as evidneced by the HT properties of the ESDs of their corresponding correlation matrices $\mathbf{X}$, and so we expect Universal properties to be similar between them.

2001; 2006; Biroli et al., 2007b). More recently, there has been mathematical work on HT random matrices (Arous & Guionnet, 2008; Auffinger et al., 2009; Burda & Jurkiewicz, 2009; Davis et al., 2014; Auffinger & Tang, 2016). There are at least 3 different Universality classes[6] of HT random matrices, defined by the range μ takes on:

- $0 < \mu < 2$: VHT: Universality class of Very Heavy-Tailed (or Lévy) matrices;
- $2 < \mu < 4$: MHT: Universality class of Moderately Heavy-Tailed (or Fat-Tailed) matrices;
- $4 < \mu$: WHT: Universality class of Weakly Heavy-Tailed matrices.

Heavy-Tailed, Finite-Size Relations. HT-RMT provides more than HT Universality classes. It also provides simple relations between the empirical observables, e.g., the PL exponent α and the maximum eigenvalue λ^{max} of each $\mathbf{W}$, with the parameter(s) μ of our generative theory, i.e, of HT-RMT.

For the VHT Universality class, the PL tail of Eqn. (2) persists in the infinite $N \to \infty$ limit, for Q fixed; and we have the linear relation between our observed exponent α and the theoretical μ:

$$\text{VHT:} \quad \alpha = \frac{1}{2}\mu + 1. \tag{4a}$$

This asymptotic expression works very well at finite size, even for very small matrices $(M, N \approx 100)$.

For the MHT Universality class, the PL tail of Eqn. (2) holds in the infinite $N \to \infty$ limit, for Q fixed, for α in Eqn. (4a). At all finite sizes, however, α is still linear in μ, but it displays *very* strong finite-size effects, empirically giving:

$$\text{MHT:} \quad \alpha = a\mu + b, \tag{4b}$$

where a, b depend strongly on M, N. (See Table 3 of (Martin & Mahoney, 2018a) for more details.) These strong finite-size effects characterize MHT distributions; and they are well-known in Statistical Physics (Sornette, 2006; Bouchaud & Potters, 2003). We will exploit these finite-size effects to develop our theory.

Finally, for both the VHT and the MHT Universality classes, the maximum empirical eigenvalue, λ^{max}, follows a Frechet distribution (i.e., an exponentially-truncated PL); and we expect it to scale with N according to Extreme Value Theory (EVT) (Biroli et al., 2007b;a; Resnick, 2007; Martin & Mahoney, 2018a):

$$\text{VHT \& MHT:} \quad \lambda^{max} \sim N^{4/\mu - 1} \tag{5}$$

(where, for simplicity, $Q = 1$).

Eqns. (4) and (5) show that we have very simple relationships that apply to random matrices that lie within both the VHT and MHT Universality classes. The α and λ^{max} are empirically-measurable quantities—of real or synthetic matrices—while μ is a parameter of the HT-RMT model. For us, the question is: how shall we *use* these relations?

Due to Heavy Tailed Mechanistic Universality (HT-MU), we expect Eqn. (5) to hold for matrices in these HT Universality classes (as evidenced by their ESD properties), e.g., DNN weight matrices $\mathbf{W}$ after training—*even when the matrix is not itself a HT random matrix* and therefore not governed by RMT or EVT. We shall use these Universal HT finite-size relations to derive a simple capacity control metric for our theory of HT-SR, and relate this to the well known product norm capacity control metric.

3. Heavy-Tailed Mechanistic Universality and Capacity Control Metrics

From prior work (Martin & Mahoney, 2018a), we expect that smaller PL exponents of the ESD imply more regularization and therefore better generalization. Since smaller norms of weight matrices often correspond to better capacity control (Liao et al., 2018; Soudry et al., 2017; Poggio

[6] Results for $\mu = 2, 4$ are slightly different (Sornette, 2006; Bouchaud & Potters, 2003). We don't describe them since we don't expect to be able to resolve them numerically. Also, sometimes Lévy matrices are split into VHT for $1 < \mu < 2$ and EHT (Extremely Heavy-Tailed) for $0 < \mu < 1$, as the properties for these two parameter regimes are somewhat different (Sornette, 2006; Bouchaud & Potters, 2003).

et al., 2018; Bartlett et al., 2017), we would like to relate the empirical PL exponent α to the empirical Frobenius norm $\|\mathbf{W}\|_F$. At least naïvely, this is a challenge, since smaller PL exponents often correspond to larger matrix norms (and thus worse generalization!). See Appendices C and D. To resolve this apparent discrepancy, we will exploit HT-MU to propose a Universal DNN complexity metric.

Form of a Proposed Universal DNN Complexity Metric. The PL exponent α is a complexity metric for a single DNN weight matrix, with smaller values corresponding to greater regularization (Martin & Mahoney, 2018a). It describes how well that matrix encodes complex correlations in the training data. Thus, a natural class of complexity or capacity metrics to consider for a DNN is to take a *weighted average*[7] of the PL exponents, $\alpha_{l,i}$, for each layer weight matrix $\mathbf{W}_{l,i}$:

$$\hat{\alpha} := \frac{1}{N_L}\sum_{l,i} b_{l,i}\alpha_{l,i}. \tag{6}$$

Here, the smaller $\hat{\alpha}$, the better we expect the DNN to represent training data, and (presumably) the better the DNN will generalize. An open question is: what are good weights $b_{l,i}$?

As we now show, we can extract the weighted average $\hat{\alpha}$ directly from the more familiar Product Norm, by exploiting both HT Universality, and its finite-size effects, arising in DNN weight matrices.

Product Norm Measures of Complexity. It has been suggested that the complexity, $\mathcal{C}$, of a DNN can be characterized by the product of the norms of layer weight matrices,

$$\mathcal{C} \sim \|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\|,$$

where $\|\mathbf{W}\|$ is, e.g., the Frobenius norm (Liao et al., 2018; Soudry et al., 2017; Poggio et al., 2018). (Here, we can use either $\|\mathbf{W}\|$ or $\|\mathbf{W}\|^2$, and one can view $\mathcal{C}$ as akin to a data-dependent VC complexity.) To that end, we consider a log complexity

$$\begin{aligned}\log \mathcal{C} &\sim \log\left[\|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\|\right] \\ &\sim \left[\log\|\mathbf{W}_1\| + \log\|\mathbf{W}_2\| \cdots \log\|\mathbf{W}_L\|\right],\end{aligned}$$

and we define the average log norm of weight matrices as

$$\langle \log \|\mathbf{W}\| \rangle = \frac{1}{N_L}\sum_l \log\|\mathbf{W}_l\|. \tag{7}$$

A Universal, Linear, PL–Norm Relation. We propose a simple linear relation between the (squared) Frobenius norm $\|\mathbf{W}\|_F^2$ of $\mathbf{W}$, the PL exponent α, and the maximum eigenvalue λ^{max} of $\mathbf{X}$ (i.e., the spectral norm $\|\mathbf{X}\|_2 = \frac{1}{N}\|\mathbf{W}\|_2^2$):

$$\textbf{PL–Norm Relation:} \quad \alpha \log \lambda^{max} \approx \log \|\mathbf{W}\|_F^2. \tag{8}$$

To our knowledge, this is the first time this PL–Norm relation has been noted in the literature (although prior work has considered norm bounds for HT data (Mahoney & Narayanan, 2009)). A few comments on Eqn. (8). First, it provides a connection between the PL parameter α of HT-SR Theory and the weight norm $\|\mathbf{W}\|_F^2$ of more traditional statistical learning theory. Second, it has the form of the well-known Hausdorff dimension (Schleicher, 2007). Third, it shows that PL exponents can alternatively be interpreted (up to the $\frac{1}{N}$ scaling) as the Stable Rank in Log-Units:

$$\text{Log-Units Stable Rank:} \quad \mathcal{R}_s^{log} := \frac{\log\|\mathbf{W}\|_F^2}{\log\lambda^{max}} \approx \alpha.$$

Our justification for proposing Eqn. (8) is three-fold.

1. We derive Eqn. (8) in the special case of very small PL exponent, $\alpha \to 1$ ($\mu \to 0$), for an $N \times M$ matrix $\mathbf{W}^{rand}(\mu)$ (with $N = M$, or $Q = 1$). See Appendix A.

2. For finite-size random matrices $\mathbf{W}^{rand}(\mu)$, the MHT Universality class, $\mu \in (2, 4)$, behaves *like* the VHT Universality class, $\mu \in (1, 2)$. Because of this similarity, we show we can extend Eqn. (8), approximately, to larger PL exponents. For $N \sim \mathcal{O}(100 - 1000)$, $\alpha \log \lambda^{max}$ increases nearly linearly with $\log \|\mathbf{W}^{rand}(\mu)\|_F^2$ as μ increases. For larger N, the relation saturates for large μ. See Appendix B.

3. *As evidence of HT-MU*, we observe empirically that Eqn. (8) also applies, approximately, to the real DNN weight matrices $\mathbf{W}$. We see that $\alpha \log \lambda^{max}$ is positively correlated with $\log \|\mathbf{W}\|_F^2$ as α increases, and even shows similar saturation effects at large α. See Appendix C.

Finally, based on Eqn. (8), we choose the weights in Eqn. (6) to be the log of the corresponding maximum eigenvalues of $\mathbf{X}$. That is, for a given l, i, we have the weights in Eqn. (6) as

$$b_{l,i} = \lambda_{l,i}^{max}.$$

[7] There are several reasons we don't want an unweighted average: an unweighted average behaves differently for HT random matrices than for well-trained DNN weight matrices, and so it would not be Universal; we want a metric that relates the α of HT-SR Theory with known capacity control metrics such as norms of weight matrices, and including weights permits this flexibility; we want weights to encode information that "larger" matrices are somehow more important; and unweighted averages, while sometimes providing predictive quality, do not perform as reliably well. See Appendices C and D for more details.

Model	Top1 Accuracy	$\hat{\alpha}$
VGG11	68.97	1.84
VGG11_BN	70.45	1.60
VGG13	69.66	1.65
VGG13_BN	71.51	1.36
VGG16	71.64	1.41
VGG16_BN	73.52	1.08
VGG19	72.08	1.16
VGG19_BN	74.27	0.81

Table 1. Results for VGG Architecture. The Top1 Accuracy is defined as the 100.0 minus the Top1 reported error.

Then, we define the complexity metrics for Linear and Convolutional Layers as follows:

$$\text{Linear Layer:} \qquad \log \|\mathbf{W}_l\|_F^2 \rightarrow \alpha_l \log \lambda_l^{max}$$

$$\text{Conv2D Layer:} \qquad \log \|\mathbf{W}_l\|_F^2 \rightarrow \sum_{i=1}^{n_l} \alpha_{l,i} \log \lambda_{l,i}^{max},$$

where, for Conv2D Layers, we relate the "norm" of the 4-index Tensor $\mathbf{W}_l$ to the sum of the $n_l = c \times d$ terms for each feature map. This lets us compare the Product Norm to the weighted average of PL exponents as follows:

$$2 \log \mathcal{C} = \langle \log \|\mathbf{W}\|_F^2 \rangle \rightarrow \hat{\alpha} := \frac{1}{N_L} \sum_{i,l} \alpha_{i,l} \log \lambda_{l,i}^{max}. \tag{9}$$

Given these connections, in Section 4, we will use $\hat{\alpha}$ to analyze numerous pre-trained DNNs.

4. Empirical Results on Pre-trained DNNs

Here, we summarize our empirical results for the VGG and ResNet series of models. See Appendix E for additional empirical results on other pre-trained DNN models.

We only consider Linear and Conv2D layers because we will only examine series of commonly available, open source, pre-trained DNNs with these kinds of layers. All models have been trained on ImageNet, and reported test accuracies are widely available. Throughout, we use the Test Accuracies for the Top1 errors (where accuracy = 100 - top1 error). We see similar results for the Top5 errors. We emphasize that, *for our analysis, we do not need to retrain these models—and we do not even need the test data!*

VGG and VGG_BN Models. We first look at the VGG class of models, comparing the log norm and the Universal $\hat{\alpha}$ metrics. See Figure 3 and Table 1 for a summary of the results. Figures 3(a) and 3(b) show both the average log Frobenius norm, $\langle \log \|\mathbf{W}\|_F \rangle$ of Eqn. (7), and the weighted average PL exponent, $\hat{\alpha}$ of Eqn. (9), as a function of the reported (Top1) test accuracy for the series of pre-trained VGG models, as available in the pyTorch package.[8] These models include VGG11, VGG13, VGG16, and VGG19, as well as their more accurate counterparts with Batch Normalization, VGG11_BN, VGG13_BN, VGG16_BN and VGG19_BN. Table 1 provides additional details.

Architecture	Model	Top 1 Accuracy	$\hat{\alpha}$
ResNet (small)	resnet10	62.54	1.94
	resnet12	63.82	0.74
	resnet14	66.83	1.70
	resnet16	69.10	1.49
ResNet18	resnet18_wd4	50.50	1.83
	resnet18_wd2	62.96	1.82
	resnet18_w3d4	66.39	0.28
	resnet18	70.48	1.09
ResNet34	resnet34	74.34	-0.42
ResNet50	resnet50	76.21	0.13
	resnet50b	76.95	0.09
ResNet101	resnet101	78.10	-0.67
	resnet101b	78.55	-0.92
ResNet152	resnet152	78.74	-1.11
	resnet152b	79.26	-1.74

Table 2. Results for ResNet Architectures and DNN Models. The Top1 Accuracy is defined as the 100.0 minus the Top1 reported error. Some $\hat{\alpha} < 0$ because the of how the ResNet weight matrices are internally scale and normalized, which makes the maximum eigenvalue less then one, $\lambda^{max} < 1$.

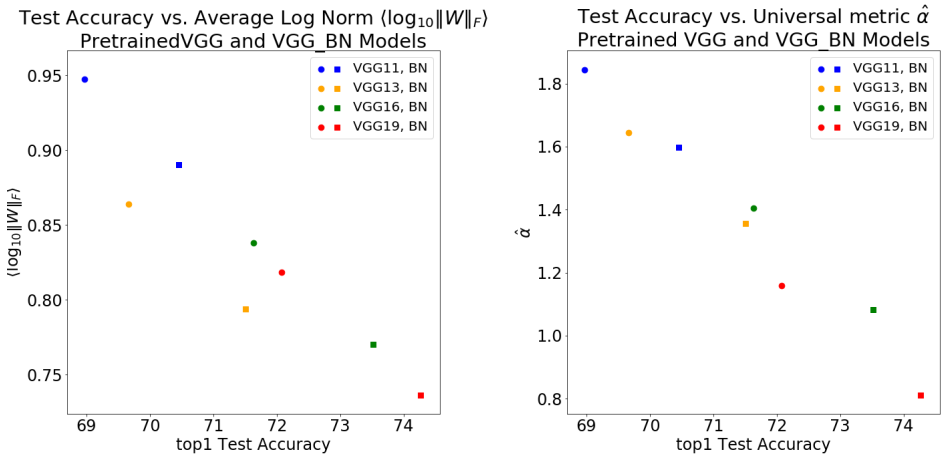
Across the entire series of architectures, reported test accuracies increase linearly as each metric, $\langle \log \|\mathbf{W}\|_F \rangle$ and $\hat{\alpha}$, decreases. Moreover, whereas the log norm relation has 2 outliers, VGG13 and VGG13_BN, the Universal $\hat{\alpha}$ metric shows a near perfect linear relation across the entire VGG series.

ResNet Models. We next look at the ResNet class of models. See Figure 4 and Table 2 for a summary of the results. Here, we consider a set of 15 different pre-trained ResNet models, of varying sizes and accuracies, ranging from the small ResNet10 up to the largest ResNet152 models, as provided by the OSMR sandbox,[9] developed for training large-scale image classification networks for embedded systems. Again, we compare the reported (Top1) test accuracy versus the average log norm $\langle \log \|\mathbf{W}\|_F \rangle$ and the Universal $\hat{\alpha}$ metrics.

As with the VGG series, both metrics monotonically decrease as the test accuracies decrease for the ResNet series, and both metrics have a few large outliers off the main line relation. See Figures 4(a) and 4(b). In particular, the log norm metric has several notable outliers, including resnet18_wd2, resnet18_wd3_d4, resnet34, and resnet10. The $\hat{\alpha}$ metric shows a slightly better relation, with resnet18_wd2 more in line, and the other 3 outliers a little less off the main line of correlation. The Universal $\hat{\alpha}$ metric is as good or slightly better than the average log norm metric

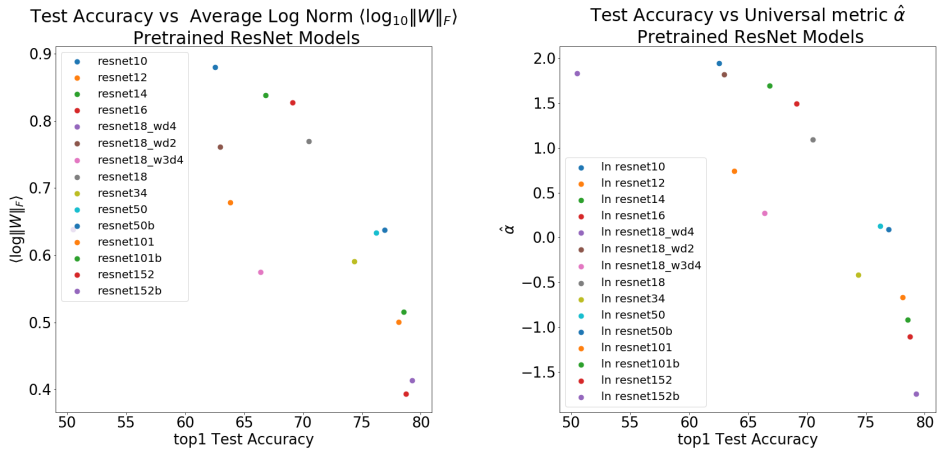
[8] https://pytorch.org/

[9] https://github.com/osmr/imgclsmob



(a) log Frobenius norm $\langle \log \|\mathbf{W}\|_F \rangle$ (b) weighted average PL exponent $\hat{\alpha}$

Figure 3. Pre-trained VGG and VGG_BN Architectures and DNNs. Top 1 Test Accuracy versus average log Frobenius norm $\langle \log \|\mathbf{W}\|_F \rangle$ (in (3(a))) or Universal, weighted average PL exponent $\hat{\alpha}$ (in (3(b))) for VGG11 vs VGG11_BN (blue), VGG13 vs VGG13_BN (orange), VGG16 vs VGG16_BN (green), and VGG19 vs VGG19_BN (red). We plot plain the VGG models with circles and the VGG_BN models with squares.



(a) log Frobenius norm $\langle \log \|\mathbf{W}\|_F \rangle$ (b) weighted average PL exponent $\hat{\alpha}$

Figure 4. Pre-trained ResNet Architectures and DNNs. Top 1 Test Accuracy versus average log Frobenius norm $\langle \log \|\mathbf{W}\|_F \rangle$ (in (4(a))) or Universal, weighted average PL exponent $\hat{\alpha}$ (in (4(b))).

for the Resnet series of models.

We see similar results for our Universal PL capacity control metric $\hat{\alpha}$ across a wide range of other pre-trained DNN models, described in Appendix E. In nearly all cases, the metric $\hat{\alpha}$ correlates well with the reported test accuracies, with only a three DNN architectures as exceptions. Overall the $\hat{\alpha}$ metric systematically correlates well with the generalization accuracy of a wide class of pre-trained DNN architectures—which is rather remarkable.

5. Discussion and Conclusion

We have presented an *unsupervised* capacity control metric which predicts trends in test accuracies of a trained DNN—without peeking at the test data. In the interests of space, see Appendix F for additional discussion. We conclude by observing simply that our work also leads to a much harder theoretical question: is it possible to characterize properties of realistic DNNs to determine whether a DNN is overtrained—without peeking at the test data?

References

Alstott, J., Bullmore, E., and Plenz, D. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, 2014.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. Technical Report Preprint: arXiv:1802.06509, 2018a.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. Technical Report Preprint: arXiv:1802.05296, 2018b.

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

Auffinger, A. and Tang, S. Extreme eigenvalues of sparse, heavy tailed random matrices. *Stochastic Processes and their Applications*, 126(11):3310–3330, 2016.

Auffinger, A., Arous, G. B., and Péché, S. Poisson convergence for the largest eigenvalues of heavy tailed random matrices. *Ann. Inst. H. Poincaré Probab. Statist.*, 45(3): 589–610, 2009.

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. Technical Report Preprint: arXiv:1706.08498, 2017.

Bartlett, P. L. For valid generalization, the size of the weights is more important than the size of the network. In *Annual Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, pp. 134–140, 1997.

Biroli, G., Bouchaud, J.-P., and Potters, M. Extreme value problems in random matrix theory and other disordered systems. *J. Stat. Mech.*, 2007:07019, 2007a.

Biroli, G., Bouchaud, J.-P., and Potters, M. On the top eigenvalue of heavy-tailed random matrices. *EPL (Europhysics Letters)*, 78(1):10001, 2007b.

Bouchaud, J.-P. and Mézard, M. Universality classes for extreme-value statistics. *Journal of Physics A: Mathematical and General*, 30(23):7997, 1997.

Bouchaud, J. P. and Potters, M. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, 2003.

Burda, Z. and Jurkiewicz, J. Heavy-tailed random matrices. Technical Report Preprint: arXiv:0909.5228, 2009.

Burda, Z., Jurkiewicz, J., Nowak, M. A., Papp, G., and Zahed, I. Lévy matrices and financial covariances. Technical Report Preprint: arXiv:cond-mat/0103108, 2001.

Burda, Z., Jurkiewicz, J., Nowak, M. A., Papp, G., and Zahed, I. Random Lévy matrices revisited. Technical Report Preprint: arXiv:cond-mat/0602087, 2006.

Cizeau, P. and Bouchaud, J. P. Theory of Lévy matrices. *Physical Review E*, 50(3):1810–1822, 1994.

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

Davis, R. A., Pfaffel, O., and Stelzer, R. Limit theory for the largest eigenvalues of sample covariance matrices with heavy-tails. *Stochastic Processes and their Applications*, 124(1):18–50, 2014.

Edelman, A. and Rao, N. R. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

Edelman, A. and Wang, Y. Random matrix theory and its innovative applications. In Melnik, R. and Kotsireas, I. (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*. Springer, 2013.

Engel, A. and den Broeck, C. P. L. V. *Statistical mechanics of learning*. Cambridge University Press, New York, NY, USA, 2001.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. Technical Report Preprint: arXiv:1809.09349, 2018.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. Technical Report Preprint: arXiv:1710.05468, 2017.

Liao, Q., Miranda, B., Banburski, A., Hidary, J., and Poggio, T. A surprising linear relationship predicts test performance in deep networks. Technical Report Preprint: arXiv:1807.09659, 2018.

Mahoney, M. W. and Narayanan, H. Learning with spectral kernels and heavy-tailed data. Technical Report Preprint: arXiv:0906.4539, 2009.

Martin, C. H. and Mahoney, M. W. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. Technical Report 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. Technical Report Preprint: arXiv:1810.01075, 2018a.

Martin, C. H. and Mahoney, M. W. Unpublished results, 2018b.

Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: on the role of implicit regularization in deep learning. Technical Report Preprint: arXiv:1412.6614, 2014.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Proceedings of the 28th Annual Conference on Learning Theory*, pp. 1376–1401, 2015.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. Technical Report Preprint: arXiv:1706.08947, 2017a.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. Technical Report Preprint: arXiv:1707.09564, 2017b.

Nishimori, H. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, Oxford, 2001.

Poggio, T., Liao, Q., Miranda, B., Banburski, A., Boix, X., and Hidary, J. Theory IIIb: Generalization in deep networks. Technical Report Preprint: arXiv:1806.11379, 2018.

Resnick, S. I. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer-Verlag, 2007.

Schleicher, D. Hausdorff dimension, its properties, and its surprises. *The American Mathematical Monthly*, 114(6): 509–528, 2007.

Sornette, D. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer-Verlag, Berlin, 2006.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. Technical Report Preprint: arXiv:1710.10345, 2017.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under- to over-parametrization affects loss landscape and generalization. Technical Report Preprint: arXiv:1810.09665, 2018.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16: 1731–1755, 2015.

Yoshida, Y. and Miyato, T. Spectral norm regularization for improving the generalizability of deep learning. Technical Report Preprint: arXiv:1705.10941, 2017.

Zhou, P. and Feng, J. Understanding generalization and optimization performance of deep CNNs. Technical Report Preprint: arXiv:1805.10767, 2018.