Universality and Capacity Metrics in Deep Neural Networks

Charles H. Martin [1] **Michael W. Mahoney** [2]

Abstract

We use the Theory of Implicit Heavy-Tailed Self-Regularization (HT-SR) to develop a new Universal capacity control metric, $\hat{\alpha}$, for Deep Neural Networks (DNNs). HT-SR indicates that modern DNNs exhibit a Heavy-Tailed Mechanistic Universality (HT-MU), meaning the spectral density of layer weight matrices can be fit to a power law, $\rho(\lambda) \sim \lambda^{-\alpha}$, with exponents, $\alpha \in [2, 5]$, that lie in common Universality classes from Heavy-Tailed Random Matrix Theory (HT-RMT). Empirically, smaller α is correlated with better generalization accuracy, with $\alpha \to 2$ universally across different best-in-class, pretrained DNN architectures. We apply this metric to over 50 different, large-scale pre-trained DNNs, ranging over 15 different architectures, trained on ImagetNet, but with differing test accuracies. This metric correlates remarkably well with reported trends in test accuracies of these DNNs, looking across each architecture (VGG16/.../VGG19, ResNet10/.../ResNet152, etc.). Our approach requires no changes to the underlying DNN or its loss function, it does not require us to train a model, and it does not even require access to the ImageNet data.

Introduction. Recent work by Martin and Mahoney (**??**) has developed a new Theory of Implicit Heavy-Tailed Self-Regularization (HT-SR) for Deep Neural Networks (DNNs). Among other things, this theory provides a Universal empirical metric that characterizes the amount of *Implicit Self-Regularization*—and, accordingly, the generalization capacity—for a wide range of publicly-available, best-in-class, pre-trained DNNs, including AlexNet, VGG, ResNet, and over 100 other models.

[1]Calculation Consulting, 8 Locksley Ave, 6B, San Francisco, CA 94122 [2]ICSI and Department of Statistics, University of California at Berkeley, Berkeley, CA 94720. Correspondence to: Charles H Martin <charles@CalculationConsulting.com>, Michael W. Mahoney <mmahoney@stat.berkeley.edu>.

Presented at the ICML 2019 Workshop on Theoretical Physics for Deep Learning.

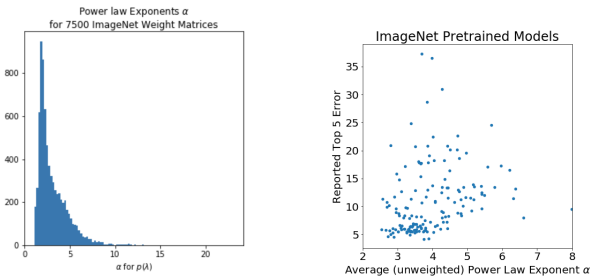
In more detail, they study the Empirical Spectral Density (ESD), $\rho(\lambda)$, of individual layer weight matrices, $\mathbf{W}$, as well as of convolutional feature maps, through the lens of Random Matrix Theory (RMT); and they observe that the individual layer ESDs almost always follow a (truncated) power law (PL) distribution

$$\rho(\lambda) \sim \lambda^{-\alpha}, \quad \lambda \leq \lambda_{max}, \tag{1}$$

where $\rho(\lambda)$ is the density of the eigenvalues λ of the normalized layer correlation matrix

$$\mathbf{X} = \frac{1}{N}\mathbf{W}^T\mathbf{W}, \tag{2}$$

and λ_{max} is the maximum eigenvalue observed. The PL exponents nearly all lie within a universal range $\alpha \in [2, 5]$, in nearly every pre-trained production-quality architecture considered. See Figure 1.



(a) Power law exponents α (b) Average α vs Top5 error

Figure 1. Histogram of PL exponents, and scatter plot of average α vs Top5 error, for nearly $10,000$ layer weight matrices (and 2D feature maps) from pre-trained DNNs, trained on ImageNet.

These observations hold across nearly $10,000$ layer weight matrices (and 2D feature maps), drawn from over 100 different, large-scale pre-trained DNNs, ranging over 15 different architectures. This includes DNNs pre-trained for computer vision (CV) tasks on ImageNet, as well as DNNs pre-trained for several different natural language processing (NLP) tasks. Moreover, smaller values of α correlate well with better generalization accuracies, with α approaching a *universal value*, $\alpha \to 2$, at the lower limit of the Moderately Heavy Tailed (or Fat Tailed) RMT Universality class (**??**) .

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 (**??**). 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 (**??**), and even models of the dynamics of actual spiking neurons (**?**). Indeed, the dynamics of learning in DNNs, and perhaps real neurons as well, seems to resemble a system near a phase transition, e.g., the phase boundary of spin glass, a system displaying Self Organized Criticality (SOC), or a Jamming transition (**??**). 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, and we use this to catalog and model the HT behavior of DNNs.

Based on these ideas, we develop here a Universal capacity control metric, $\hat{\alpha}$. This metric is a weighted average of the layer PL exponents, α_l, of the DNN layer weight matrices,

$$\hat{\alpha} = \frac{1}{N_L}\sum_{l\in L}\alpha_l \log \lambda_l^{max}, \tag{3}$$

where L indexes layers, and λ_l^{max} is the maximum eigenvalue (i.e., Spectral norm) of layer correlation matrices $\mathbf{W}_l$.

Approach. Our approach and intent differ from other theoretical studies in the DNN literature, although we can relate our results back to known results. Most recently, e.g., Liao et al. (**?**) used an appropriately-scaled, data-dependent Product Norm capacity control metric to bound worst-case generalization error for several small (not 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 (**???**) and (**??????????**), as well as much older (**??**). These studies seek *worst-case* complexity bounds, motivated in some cases to reconcile discrepancies with more traditional statistical learning theory, and they apply (when applied at all) to quite small NNs.

This approach contrasts with that of Martin and Mahoney (**??**), who looked at empirical properties of a wide range of state-of-the-art models, and from this developed a metric that provides *a posteriori* characterization of implicit regularization for typical-case DNNs. Thus, instead of using statistical learning theory principles to propose a metric that provides worst-case *a priori* bounds, we seek here an *average-case* or *typical case* (where "typical" is for current state-of-the-art publicly-available pre-trained DNN models) complexity metric, viable in production settings as a guide to the development of better DNNs at scale.

Theory. Let us write the Energy Landscape (or optimization function) for a typical DNN with L layers, with activation functions $h_l(\cdot)$, and with weight matrices and biases $\mathbf{W}_l$ and $\mathbf{b}_l$, as follows:

$$E = h_L(\mathbf{W}_L \times h_{L-1}(\mathbf{W}_{L-1} \times (\cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L). \tag{4}$$

Typically, this model would be trained on some labeled data $\{d_i, y_i\} \in \mathcal{D}$, using Backprop (**?**), by minimizing the loss $\mathcal{L} = \sum_{i\in\mathcal{D}}[E(d_i) - y_i]$.

For simplicity, we do not indicate the structural details of the layers (e.g., Dense or not, Convolutions or not, Residual/Skip Connections, etc.), nor do we consider the details of the optimizer or the training process.

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 directly from 2D Convolutional (Conv2D) layers. (We have not yet analyzed LSTM or other complex Layers.) A typical modern DNN may have anywhere between 5 and 5000 2D layer matrices / feature maps.

Our capacity metric $\hat{\alpha}$ depends two parameters, the PL exponent α (a measure of matrix sparsity) and the maximum eigenvalue λ_{max} (which corrects for the matrix scale). Other capacity metrics typically consider either just the sparsity or rank-based sparsity (e.g., the Stable Rank) or just the scale via a matrix norm (e.g., Spectral or Frobenius norm).

We can relate $\hat{\alpha}$ to these more familiar, data dependent metrics. Consider the Product Norm capacity metric, $\mathcal{C}$, defined as

$$\mathcal{C} \sim \|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\|. \tag{5}$$

Using a standard trick from field theory, we consider the log Product Norm, which takes the form of an average log norm

$$\begin{aligned}
\log \mathcal{C} &\sim \log\left[\|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\|\right] \\
&\sim \left[\log\|\mathbf{W}_1\| + \log\|\mathbf{W}_2\| \cdots \log\|\mathbf{W}_L\|\right] \\
&\sim \langle \log\|\mathbf{W}\|\rangle = \frac{1}{N_L}\sum_l \log\|\mathbf{W}_l\|.
\end{aligned}$$

When $\|\mathbf{W}\|$ is the Spectral norm, $\|\mathbf{W}\|_2 \sim \lambda_{max}$, then our $\hat{\alpha}$ of Eqn. (3) is a weighted average of the log Spectral norms, where the weights are power law exponents α. In this sense, our universal metric $\hat{\alpha}$ behaves like an *average-case* version of what is a worst-case bound, but it is more suitable for applying to large, production-level DNNs.

When $\|\mathbf{W}\|$ is the Frobenius norm, $\|\mathbf{W}\|_F^2$, we can use results of HT RMT to interpret the PL exponents α as a type of Soft or Stable Rank. Specifically, when α is very small, we can relate α to the more familiar Stable Rank $\mathcal{R}_s^{log}$, expressed in log-units (and up to the $\frac{1}{N}$ scaling):

$$\mathcal{R}_s^{log} := \frac{\log\|\mathbf{W}\|_F^2}{\log\lambda^{max}} \approx \alpha. \tag{6}$$

Using this, one could implement our capacity metric as a regularizer to improve DNN training by implementing a

Stable Rank regularizer (similar to how Spectral/Frobenius norm regularization is often implemented.

Methodology. To evaluate our metric, we introduce a new methodology to analyze the performance of large-scale pre-trained DNNs, including the VGG and ResNet series of models, as well nearly 100 other widely available models, and we study how capacity metrics correlate with the reported test accuracies.

This approach offers several advantages over common practice in the area, most notably the following.

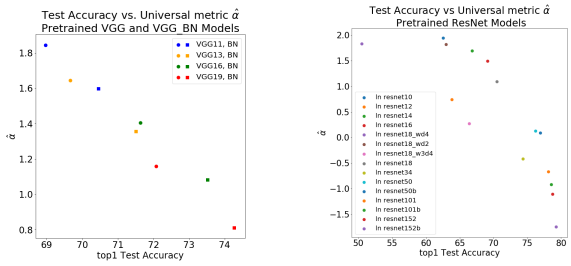
- We do not need access to the original ImageNet data, just the pre-trained models (i.e., as distributed with PyTorch, on github, and/or from the ModelZoo).
- We do not need to engage in expensive training/retraining, architecture adjustment, hyperparameter tuning, etc.
- Our results are easily reproducible. To make things more reproducible, we provide a python command tool, `WeightWatcher` (**?**), that works with both PyTorch (v1) and Keras (v2) models, and that computes a wide range of average log capacity metrics.

We have applied our Universal capacity control metric $\hat{\alpha}$ to a wide range of large-scale pre-trained production-level DNNs. For Linear DNN layers, we simply replace the log Norm with our metric, whereas for Conv2D Layers, we associate 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, as follows:

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

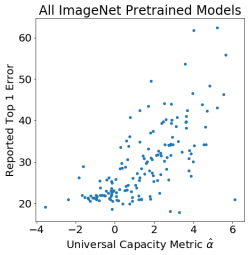
$$\text{Conv2D Layer:} \qquad \log \|\mathbf{W}_l\| \rightarrow \sum_{i=1}^{n_l} \alpha_{l,i} \log \lambda_{l,i}^{max}.$$

Results. Our Universal metric correlates very well with the reported average test accuracies across many series of pre-trained DNNs. See Figures 2 and 3, where In Figure 2, we present two complex examples of pre-trained models with similar architectures of differing depths. Figure 2(a) presents the VGG series of networks (VGG11, 13, 16, and 19, with an without Batch Normalization (BN), as available in PyTorch), and Figure 2(b) presents results for a large set of ResNet architectures, ranging from ResNet10 to ResNet152 (available here). We generate the $\hat{\alpha}$ metric using the `WeightWatcher` tool (**?**), (hopefully) making these results trivial to reproduce. Notice, we do not have access to the (ImageNet) test data. Compare $\hat{\alpha}$ to reported Top1 test errors. Amazingly, our simple $\hat{\alpha}$ metric correlates very well with the reported Top1 test errors. [michael: CHARLES: Finalize Figure 3, and add sentences here and in caption.]



(a) $\hat{\alpha}$ for VGG series (b) $\hat{\alpha}$ for ResNet series

Figure 2. Top 1 Test Accuracy versus $\hat{\alpha}$ for pre-trained VGG and ResNet Architectures and DNNs.



(a) $\hat{\alpha}$ vs Top1 error

Figure 3. Caption: we have room for 1 more [michael: CHARELS: do we want another subfigure here, maybe scatter plot of some NLP models.][charles: we do not have NLP data across architectures...still thinking about this]

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 any series of DNNs, let alone 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 can be reproduced with the `WeightWatcher` package (**?**); and our results suggest that our "practical theory" methodology is fruitful more generally for engineering good algorithms for realistic large-scale DNNs.

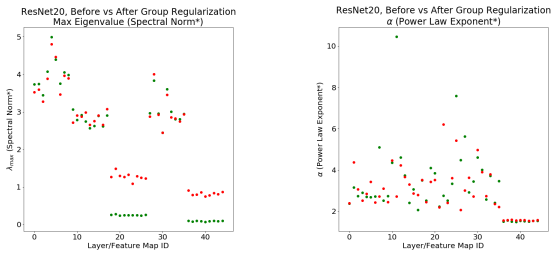
Comparison with other metrics. Our Universality metric $\hat{\alpha}$ can be related to other, more familiar capacity metrics such as the Spectral Norm, the Frobenius Norm, and robust measures of matrix rank such as the Stable Rank. This suggests an obvious question: Does the $\hat{\alpha}$ metric "work" simply because DNN models have explicit regularization (i.e., with a Frobenius or Spectral Norm constraint)? More generally, is the $\hat{\alpha}$ metric just a variation of these more familiar norm-based metrics?

The short answer is that we believe that HTMU is a new, more fundamental relation. It complements other metrics, but it works more generally on actual DNNs when other norm-based theoretical metrics inspired by worst-case

bounds fail. In particular, we can identify counter examples, most notably in compressed DNN models (**?**). For compressed models, we have observed that the average Frobenius Norm increases with decreasing test error, whereas the average α decreases, as expected.

To illustrate this, we consider average metrics measured on ResNet20, trained on CIFAR10, before and after applying the Group Regularization technique, as implemented in the `distiller` package.[1] See Figure 4.

[michael: CHARLES: More details here, e.g., what are axes, what is green versus red, more description of what figures show, etc.] [charles: We analyze the available pre-trained 4D_regularized_5Lremoved baseline and finetuned models(**?**). Figure **??** presents the maximum eigenvalues (λ_{max}, or Spectral Norm) and PL exponent α for each individual layer weight matrix $\mathbf{W}_l$] [2] The reported baseline test accuracies ($Top1 = 91.450$ and $Top5 = 99.750$) are better than the reported fine-tuned test accuracies ($Top1 = 91.020$ and $Top5 = 99.670$), so traditional theory suggests that the baseline Spectral Norm ($\lambda_{max} \sim \|\mathbf{W}\|_2$) should be *smaller* than those of the layers in the fine-tuned model. Based on previous empirical results, we may also expect the baseline Frobenius norm to be smaller than the fine tuned. In both cases (Frobenius norm results not shown), we observe the opposite.



(a) λ_{max} for ResNet20 layers (b) α for ResNet20 layers

Figure 4. Analysis of ResNet20 , distilled with Group Regularization, as implemented in the `distiller` (4D_regularized_5Lremoved) pre-trained models. Comparison of individual layer $\mathbf{W}_l$ maximum eigenvalues (λ_{max}, or Spectral Norms) and Power Law exponent α, between baseline (green) and fine-tuned (red) pre-trained models. [michael: CHARLES: better explanation here and/or text; and is the "*with" above a typo.][charles: see now]

The `distiller` Group Regularization technique has the unusual effect of increasing the norms of the $\mathbf{W}$ feature maps for at least 2 of the Conv2D layers. We suspect this effect arises because the Group Regularization concentrates Frobenius mass from the five removed Conv2D layers into the remaining Conv2D layers.

Notice while the matrix norms behave atyically, the layers α do not systematically differ between the baseline and fine-tuned models. Also (not shown), the average (unweighted) baseline α is indeed smaller than the fine-tuned average, as would be predicted by HT-SR Theory.

Discussion. We have presented an *unsupervised* capacity control metric which predicts trends in test accuracies of a large-scale pre-trained DNN—without even peeking at the training data or the test data. This complexity metic, $\hat{\alpha}$ of Eqn. (3), is a weighted average of the PL exponents α for each layer weight matrix, where α is defined in the recent HT-SR Theory (**??**), and where the weights are the largest eigenvalue λ^{max} of the correlation matrix $\mathbf{X}$. We examine several commonly-available, pre-trained, production-quality DNNs, by plotting $\hat{\alpha}$ versus the reported test accuracies. This covers classes of DNN architectures including the VGG models, ResNet, DenseNet, etc. In nearly every class, and except for only a few counterexamples, smaller $\hat{\alpha}$ corresponds to better average test accuracies, thereby providing a strong predictor of model quality for large-scale state-of-the-art DNN models.

It is worth emphasizing that we are taking a very non-standard approach (at least for the DNN and ML communities). We did not train/retrain lots and lots of (typically rather small) models, analyzing training/test curves, trying to glean from them bits of insight that might then extrapolate to much-larger more realistic models. Instead, we take advantage of the fact that there already exist many (typically rather large) publicly-available pre-trained models, and we analyze the properties of these models. That is, we viewed these publicly-available pre-trained models as artifacts of the world that achieve state-of-the-art performance in CV, NLP, and related applications; and we attempted to understand why. To do this, we extracted data-dependent metrics to predict generalization performance on production-quality models. Given well-known challenges associated with training, and given our results here as well as other recent results (**??**), we suggest that this methodology be applied more generally.

In theoretical physics, many researchers study neural networks using spin glass models, such as the traditional Gardner analysis (**??**) and more recent work (**??**). Most notably, however, Choromanska et. al. have suggested that the energy landscape of DNNs should resemble the zero-temperature energy landscape of a p-spin spherical spin glass (**?**). Specifically, this implies there are many local minima that concentrate at a floor just above the global minima. Here, however, and following previous results (**??**), we argue that such spin glass models should really employ

[1] For details, see `https://nervanasystems.github.io/distiller/#distiller-documentation` and also `https://github.com/NervanaSystems/distiller`.

[2] We only include layer matrices or feature maps with $M \geq 50$

HT, not Gaussian, stochastic spin-spin interactions. Such models would have a *very* different zero-temperature complexity. Indeed, HT Levy spin-glasses do *not* have a large number of low lying minima (**???**). Instead, they resemble something like a ruggedly-convex "energy funnel," with few local minima, similar in some sense to the Wolynes-Onuchic Energy landscape (**??**) suggested in the early protein folding literature.