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 (Martin & Mahoney, 2018; 2019) has develoepd 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, etc.

[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.

Instead of using statistical learning theory principles to proposa a metric that provides worst-case *a priori* bounds, Martin and Mahoney look at empirical properties of a wide range of state-of-the-art models, and from this they develop a metric that provides *a posteriori* predictions for the typical case (where "typical" is for current state-of-the-art publicly-available pre-trained models).

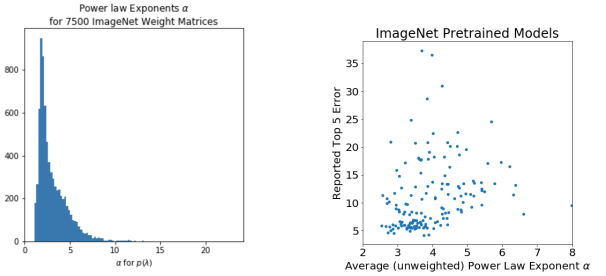
In more detail, they study the Empirical Spectral Density (ESD), $\rho(\lambda)$, of individual layer weight matrices, $\mathbf{W}$, as well as of convolutional feature maps, through the lens of Random Matrix Theory (RMT); and they observe that the individual layer ESDs almost always follow a power law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

$$\rho(\lambda) \sim \lambda^{-\alpha}, \tag{1}$$

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

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

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 CV and NLP.

These observations hold across nearly $10,000$ layer weight matrices (and 2D feature maps), including 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 (Martin & Mahoney, 2018; 2019) .

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, 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 (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, and we use this to catalog and model the HT behavior of DNNs.

Based on these ideas, here, we develop 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,

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

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

Theory. Our approach and intent differ from other theoretical studies in the DNN literature, although we can relate our results back to known results. For example, 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). 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. We seek an *average-case* or *typical case* (for realistic large-scale problems) complexity metric, viable in production settings as a guide to the development of better DNNs at scale.

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 (LeCun et al., 2012), 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.

We can relate our universality metric to the more traditional, data dependent, VC-like product norm capacity metrics $\mathcal{C}$. Let us define the *worst-case* bound $\mathcal{C}$ as [michael: why is this worst case, and why is this a bound.]

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

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 $\hat{\alpha}$ is a weighted average of the log product Spectral norm, where the weights are power law exponents α. [michael: I think we are arguing by analogy here.] 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 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 norm regularization is 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 200 other widely available models, and we study how capacity metrics correlate with the reported test accuracies. [michael: Should we describe it more, and move some relevant comments from elsewhere to here.]

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

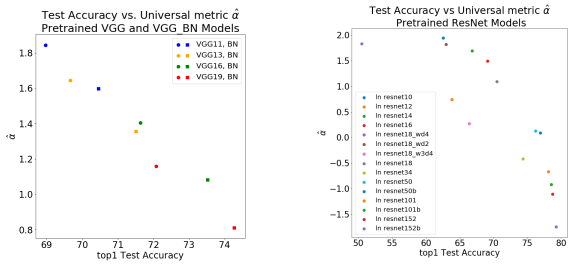
- We do not need access to the original ImageNet data, just the pre-trained models (i.e., as distributed with PyTorch, on github, and/or from the ModelZoo).
- Our results are easily reproducible. To make things more reproducible, we provide a python command tool, weightwatcher (wei), that works with both PyTorch (v1) and Keras (v2) models and 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 can simply replace the log Norm with our metric, whereas for Conv2D Layers, we can 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:

$$\begin{aligned} \text{Linear Layer:} \quad & \log \|\mathbf{W}_l\| \rightarrow \alpha_l \log \lambda_l^{max} \\ \text{Conv2D Layer:} \quad & \log \|\mathbf{W}_l\| \rightarrow \sum_{i=1}^{n_l} \alpha_{l,i} \log \lambda_{l,i}^{max}. \end{aligned}$$

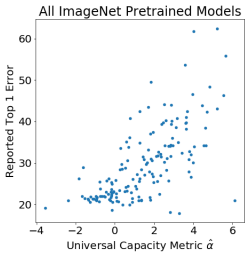
Results. Our Universal metric correlates very well with the reported average test accuracies across many series of pre-trained DNNs. See Figures 2 and 3.

[michael: Prob make a short paragraph of this.] DISCUSS FIGURE IN DETAIL, CAN ADD 1 more BELOW and discuss



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

Figure 2. Top 1 Test Accuracy versus the Universal, weighted average PL exponent $\hat{\alpha}$ 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: Do we want another subfigure here, maybe scatter plot of some NLP models.]

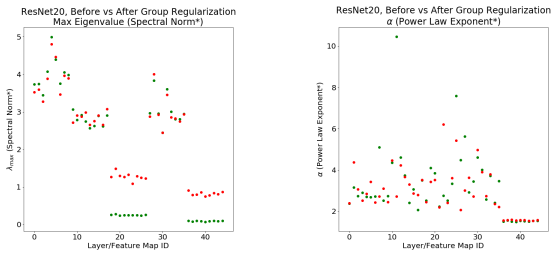
Our empirical results are, to our knowledge, the first time such theoretical capacity metrics have been reported to predict (trends in) the test accuracy for *pre-trained production-level* DNNs. In particular, this illustrates the usefulness of these norm-based metrics beyond smaller models such as MNIST, CIFAR10, and CIFAR100. Our results can be reproduced with the `WeightWatcher` package (wei); 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 $\hat{\alpha}$ is related to other, more familiar capacity metrics such as the Spectral Norm, the Frobenius $(L2)$ 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 an $(L2)$ or Spectral Norm constraint)? More generally, is the $\hat{\alpha}$ metric just a variation of these more familiar *worst case* bounds? [michael: We seem to be convolving norms versus exponents with worst-case versus typical-case issues.]

The short answer is, we believe that Mechanistic Univer-

sality is a new, more fundamental relation. It complements other metrics, but it works more generally on actual DNN models when other theoretical metrics based on worst-case bounds fail. In particular, we can identify counter examples, most notably in compressed DNN models, where the average Frobenius $(L2)$ Norm increases with decreasing test error, but the average α decreases, as expected. [michael: Citations for compressed DNN modes.]

We consider average metrics measured on ResNet20, trained on CIFAR10, before and after applying the Group Regularization technique, as implemented in the *distiller* package ADDCITE. [michael: Some sort of reference.] See Figure 4. Notice that 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. We observe the reverse for both.



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

Figure 4. ResNet20 with Group Regularization, comparison of layer $\mathbf{W}$ maximum eigenvalue (λ_{max}, or Spectral Norm *with $1/N$ normalization) and Power Law exponent α pre-trained baseline and fine-tuned models. [michael: Better explanation.]

[michael: Is this a separate par, or ocmbine with above.] DESCRIBE THESE 2 FIGURES

For the $4D\_regularized$ models, the *distiller* Group Regularization technique has the unusual effect of increasing the norms of the 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 5 removed Conv2D layers into these remaining Conv2D layers.

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

[michael: TO DO: COMMMENT ON IMPLICATIONS.]

Discussion. We have presented an *unsupervised* capacity control metric which predicts trends in test accuracies of a trained DNN—without peeking at 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 (Martin & Mahoney, 2018; 2019), 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 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 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 then extracted data-dependent metrics to predict their generalization performance on production-quality models. Given well-known challenges associated with training, and given our results here as well as other recent results (Martin & Mahoney, 2018; 2019), 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 (Gardner & Derrida, 1989) and more recent work (Pennington et al., 2017; 2018). 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 (Choromanska et al., 2014). Specifically, this implies there are many local minima that concentrate at a floor just above the global minima. Here, however, and following our previous results (Martin & Mahoney, 2018; 2019), we argue that such spin glass models should really employ 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 (Cizeau & Bouchaud, 1993; Galluccio et al., 1998; Gábora & Kondor, 1999). Instead, they resemble something like a ruggedly-convex "energy funnel," with few local minima, similar in some sense to the Wolynes-Onuchic Energy

landscape (Bryngelson et al., 1995; Onuchic et al., 1997) suggested in the early protein folding literature.

References

`https://pypi.org/project/WeightWatcher/`.

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.

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.

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

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21, 1995.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. Technical Report Preprint: arXiv:1412.0233, 2014.

Cizeau, P. and Bouchaud, J. P. Mean field theory of dilute spin-glasses with power-law interactions. *Journal of Physics A: Mathematical and General*, 26(5):L187–L194, 1993.

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

Gábora, A. and Kondor, I. Portfolios with nonlinear constraints and spin glasses. *Physica A: Statistical Mechanics and its Applications*, 274(12):222–228, 1999.

Galluccio, S., Bouchaud, J.-P., and Potters, M. Rational decisions, random matrices and spin glasses. *Physica A*, 259:449–456, 1998.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983–1994, 1989.

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.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient BackProp. In Montavon, G., Orr, G. B., and Müller, K.-R. (eds.), *Neural Networks: Tricks of the Trade: Second Edition*, pp. 9–48. Springer-Verlag, 2012.

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. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. Technical Report Preprint: arXiv:1810.01075, 2018.

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

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.

Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600, 1997.

Pennington, J., Schoenholz, S. S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. Technical Report Preprint: arXiv:1711.04735, 2017.

Pennington, J., Schoenholz, S. S., and Ganguli, S. The emergence of spectral universality in deep networks. Technical Report Preprint: arXiv:1802.09979, 2018.

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.

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.

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.