Traditional and Heavy Tailed Self Regularization in Neural Network Models

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

Abstract

Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neural Networks (DNNs), including both production quality, pre-trained models such as AlexNet and Inception, and smaller models trained from scratch, such as LeNet5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate that the empirical spectral density (ESD) of DNN layer matrices displays signatures of traditionally-regularized statistical models, even in the absence of exogenously specifying traditional forms of regularization, such as Dropout or Weight Norm constraints. Building on recent results in RMT, most notably its extension to Universality classes of Heavy-Tailed matrices, we develop a theory to identify *5+1 Phases of Training*, corresponding to increasing amounts of *Implicit Self-Regularization*. For smaller and/or older DNNs, this Implicit Self-Regularization is like traditional Tikhonov regularization, in that there is a "size scale" separating signal from noise. For state-of-the-art DNNs, however, we identify a novel form of *Heavy-Tailed Self-Regularization*, similar to the self-organization seen in the statistical physics of disordered systems. This implicit Self-Regularization can depend strongly on the many knobs of the training process. We demonstrate that we can cause a small model to exhibit all 5+1 phases simply by changing the batch size.

1. Introduction

The inability of optimization and learning theory to explain and predict the properties of NNs is not a new phenomenon. From the earliest days of DNNs, it was suspected that VC theory did not apply to these systems (Vapnik et al., 1994). It was originally assumed that local minima in the energy/loss surface were responsible for the inability of VC theory to describe NNs (Vapnik et al., 1994), and that the mechanism for this was that getting trapped in local minima during training limited the number of possible functions realizable by the network. However, it was very soon realized that the presence of local minima in the energy function was *not* a problem in practice (LeCun et al., 1998; Duda et al., 2001). Thus, another reason for the inapplicability of VC theory was needed. At the time, there did exist other theories of generalization based on statistical mechanics (Seung et al., 1992; Watkin et al., 1993; Haussler et al., 1996; Engel & den Broeck, 2001), but for various technical and nontechnical reasons these fell out of favor in the ML/NN communities. Instead, VC theory and related techniques continued to remain popular, in spite of their obvious problems.

More recently, theoretical results of Choromanska et al. (Choromanska et al., 2014) (which are related to (Seung et al., 1992; Watkin et al., 1993; Haussler et al., 1996; Engel & den Broeck, 2001)) suggested that the Energy/optimization Landscape of modern DNNs resembles the Energy Landscape of a zero-temperature *Gaussian* Spin Glass; and empirical results of Zhang et al. (Zhang et al., 2016) have again pointed out that VC theory does not describe the properties of DNNs. Martin and Mahoney then suggested that the Spin Glass analogy may be useful to understand severe overtraining versus the inability to overtrain in modern DNNs (Martin & Mahoney, 2017).

Motivated by this, we are interested here in two questions.

- **Theoretical Question.** Why is regularization in deep learning seemingly quite different than regularization in other areas on ML; and what is the right theoretical framework with which to investigate regularization for DNNs?
- **Practical Question.** How can one control and adjust, in a theoretically-principled way, the many knobs and switches that exist in modern DNN systems, e.g., to train these models efficiently and effectively, to monitor their effects on the global Energy Landscape, etc.?

That is, we seek a *Practical Theory of Deep Learning*, one that is prescriptive and not just descriptive. This (phenomenological) theory would provide useful tools for practitioners wanting to know *How* to characterize and control the Energy Landscape to engineer larger and betters DNNs;

[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>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and it would also provide theoretical answers to broad open questions as *Why* Deep Learning even works.

Main Empirical Results. Our main empirical results consist in evaluating empirically the ESDs (and related RMT-based statistics) for weight matrices for a suite of DNN models, thereby probing the Energy Landscapes of these DNNs. For older and/or smaller models, these results are consistent with implicit *Self-Regularization* that is Tikhonov-like; and for modern state-of-the-art models, these results suggest novel forms of *Heavy-Tailed Self-Regularization*.

• **Self-Regularization in old/small models.** The ESDs of older/smaller DNN models (like LeNet5 and a toy MLP3 model) exhibit weak *Self-Regularization*, well-modeled by a perturbative variant of Marchenko-Pastur (MP) theory, the Spiked-Covariance model. Here, a small number of eigenvalues pull out from the random bulk, and thus the MP Soft Rank (defined in Section 4) and Stable Rank both decrease. This weak form of *Self-Regularization* is like Tikhonov regularization, in that there is a "size scale" that cleanly separates "signal" from "noise," but it is different than explicit Tikhonov regularization in that it arises implicitly due to the DNN training process itself.

• **Heavy-Tailed Self-Regularization.** The ESDs of larger, modern DNN models (including AlexNet and Inception and nearly every other large-scale model we have examined) deviate strongly from the common Gaussian-based MP model. Instead, they appear to lie in one of the very different Universality classes of Heavy-Tailed random matrix models. We call this *Heavy-Tailed Self-Regularization*. The ESD appears Heavy-Tailed, but with finite support. In this case, there is *not* a "size scale" (even in the theory) that cleanly separates "signal" from "noise."

Main Theoretical Results. Our main theoretical results consist in an phenomenological theory for DNN Self-Regularization. Our theory uses ideas from RMT—both vanilla MP-based RMT as well as extensions to other Universality classes based on Heavy-Tailed distributions—to provide a visual taxonomy for $5 + 1$ *Phases of Training*, corresponding to increasing amounts of Self-Regularization.

• **Modeling Noise and Signal.** We assume that a weight matrix $\mathbf{W}$ can be modeled as $\mathbf{W} \simeq \mathbf{W}^{rand} + \Delta^{sig}$, where $\mathbf{W}^{rand}$ is "noise" and where Δ^{sig} is "signal." For small to medium sized signal, $\mathbf{W}$ is well-approximated by an MP distribution—with elements drawn from the Gaussian Universality class—perhaps after removing a few eigenvectors. For large and strongly-correlated signal, $\mathbf{W}^{rand}$ gets progressively smaller, but we can model the non-random strongly-correlated signal Δ^{sig} by a Heavy-Tailed random matrix, i.e., a random matrix with elements drawn from a Heavy-Tailed (rather than Gaussian) Universality class.

• **5+1 Phases of Regularization.** Based on this, we construct a practical, visual taxonomy for 5+1 Phases of Training. Each phase is characterized by stronger, visually distinct signatures in the ESD of DNN weight matrices, and successive phases correspond to decreasing MP Soft Rank and increasing amounts of *Self-Regularization*. The 5+1 phases are: RANDOM-LIKE, BLEEDING-OUT, BULK+SPIKES, BULK-DECAY, HEAVY-TAILED, and RANK-COLLAPSE.

Based on these results, we speculate that all well optimized, large DNNs will display *Heavy-Tailed Self-Regularization*.

Evaluating the Theory. We also provide a detailed evaluation of our theory using a smaller MiniAlexNew model.

• **Effect of Explicit Regularization.** We analyze ESDs of MiniAlexNet by removing all explicit regularization (Dropout, Weight Norm constraints, Batch Normalization, etc.) and characterizing how the ESD of weight matrices behave during and at the end of Backprop training, as we systematically add back in different forms of explicit regularization.

• **Exhibiting the 5+1 Phases.** We demonstrate that we can exhibit all 5+1 phases by appropriate modification of the various knobs of the training process. In particular, by decreasing the batch size from 500 to 2, we can make the ESDs of the fully-connected layers of MiniAlexNet vary continuously from RANDOM-LIKE to HEAVY-TAILED, while increasing generalization accuracy along the way. These results illustrate the *Generalization Gap* pheneomena (Hoffer et al., 2017; Keskar et al., 2016; Goyal et al., 2017), and they explain that pheneomena as being caused by the implicit Self-Regularization associated with models trained with smaller and smaller batch sizes.

We should note that since the initial dissemination of these results, our theory has been *used* to develop a Universal capacity control metric to predict trends in test accuracies for large-scale pre-trained DNNs (Martin & Mahoney, 2019). A longer and more detailed version of this paper is available in technical report form (Martin & Mahoney, 2018).

2. Basic Random Matrix Theory (RMT)

In this section, we summarize results from RMT that we use.

2.1. Marchenko-Pastur (MP) theory

MP theory considers the density of singular values $\rho(\nu_i)$ of random rectangular matrices $\mathbf{W}$. This is equivalent to considering the density of eigenvalues $\rho(\lambda_i)$, i.e., the ESD, of matrices of the form $\mathbf{X} = \frac{1}{N}\mathbf{W}^T\mathbf{W}$. MP theory then makes strong statements about such quantities as the shape of the distribution in the infinite limit, it's bounds, expected finite-size effects, such as fluctuations near the edge, and rates of convergence.

To apply RMT, we need only specify the number of rows and columns of $\mathbf{W}$ and assume that the elements $W_{i,j}$ are drawn from a distribution that is a member of a certain *Universality class* (there are different results for different Universality classes). RMT then describes properties of the ESD, even at finite size; and one can compare predictions of RMT with empirical results. Most well-known is the Universality class of Gaussian distributions. This leads to the basic or vanilla MP theory. More esoteric—but ultimately more useful for us—are Universality classes of Heavy-Tailed distributions.

Gaussian Universality class. We start by modeling $\mathbf{W}$ as an $N \times M$ random matrix, with elements from a Gaussian distribution, such that: $W_{ij} \sim N(0, \sigma_{mp}^2)$. Then, MP theory states that the ESD of the correlation matrix, $\mathbf{X}$, has the limiting density given by the MP distribution $\rho(\lambda)$:

$$\rho_N(\lambda) \xrightarrow[Q \text{ fixed}]{N\to\infty} \frac{Q}{2\pi\sigma_{mp}^2} \frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{\lambda}, \quad (1)$$

if $\lambda \in [\lambda^-, \lambda^+]$, and 0 otherwise. Here, σ_{mp}^2 is the element-wise variance of the original matrix, $Q = N/M \geq 1$ is the aspect ratio of the matrix, and the minimum and maximum eigenvalues, $\lambda^\pm$, are given by

$$\lambda^\pm = \sigma_{mp}^2 \left(1 \pm \frac{1}{\sqrt{Q}}\right)^2 . \quad (2)$$

Finite-size Fluctuations at the MP Edge. In the infinite limit, all fluctuations in $\rho_N(\lambda)$ concentrate very sharply at the MP edge, $\lambda^\pm$, and the distribution of the maximum eigenvalues $\rho_\infty(\lambda_{max})$ is governed by the Tracy Widom (TW) Law. Even for a single finite-sized matrix, however, MP theory states the upper edge of $\rho(\lambda)$ is very sharp; and even when the MP Law is violated, the TW Law, with finite-size corrections, works very well at describing the edge statistics. When these laws are violated, this is very strong evidence for the onset of more regular non-random structure in the DNN weight matrices, which we will interpret as evidence of *Self-Regularization*.

2.2. Heavy-Tailed extensions of MP theory

MP-based RMT is applicable to a wide range of matrices; but it is *not* in general applicable when matrix elements are strongly-correlated. Strong correlations appear to be the case for many well-trained, production-quality DNNs. In statistical physics, it is common to *model* strongly-correlated systems by Heavy-Tailed distributions (Sornette, 2006). The reason is that these models exhibit, more or less, the same large-scale statistical behavior as natural phenomena in which strong correlations exist (Sornette, 2006; Bouchaud & Potters, 2011). Moreover, recent results from MP/RMT have shown that new Universality classes exist for matrices with elements drawn from certain Heavy-Tailed distributions (Bouchaud & Potters, 2011).

We use these Heavy-Tailed extensions of basic MP/RMT to build an operational and phenomenological theory of Regularization in Deep Learning; and we use these extensions to justify our analysis of both *Self-Regularization* and *Heavy-Tailed Self-Regularization*. Briefly, our theory for simple *Self-Regularization* is insipred by the Spiked-Covariance model of Johnstone (Johnstone, 2001) and its interpretation as a form of *Self-Organization* by Sornette (Malevergne & Sornette, 2002); and our theory for more sophisticated *Heavy-Tailed Self-Regularization* is inspired by the application of MP/RMT tools in quantitative finance by Bouchuad, Potters, and coworkers (Galluccio et al., 1998; Laloux et al., 1999; 2005; Biroli et al., 2007b;a; Bouchaud & Potters, 2011; Bun et al., 2017), as well as the relation of Heavy-Tailed phenomena more generally to *Self-Organized Criticality* in Nature (Sornette, 2006).

Universality classes for modeling strongly correlated matrices. Consider modeling $\mathbf{W}$ as an $N \times M$ random matrix, with elements drawn from a Heavy-Tailed—e.g., a Pareto or Power Law (PL)—distribution:

$$W_{ij} \sim P(x) \sim \frac{1}{x^{1+\mu}}, \ \ \mu > 0. \quad (3)$$

In these cases, if $\mathbf{W}$ is element-wise Heavy-Tailed, then the ESD $\rho_N(\lambda)$ exhibits Heavy-Tailed properties, either globally for the entire ESD and/or locally at the bulk edge.

Table 1 summarizes these recent results, comparing basic MP theory, the Spiked-Covariance model, and Heavy-Tailed extensions of MP theory, including associated Universality classes. To apply the MP theory, at finite sizes, to matrices with elements drawn from a Heavy-Tailed distribution of the form given in Eqn. (3), we have one of the following three Universality classes.

- **(Weakly) Heavy-Tailed**, $4 < \mu$: Here, the ESD $\rho_N(\lambda)$ exhibits "vanilla" MP behavior in the infinite limit, and the expected mean value of the bulk edge is $\lambda^+ \sim M^{-2/3}$. Unlike standard MP theory, which exhibits TW statistics at the bulk edge, here the edge exhibits PL / Heavy-Tailed fluctuations at finite N. These finite-size effects appear in the edge / tail of the ESD, and they make it hard or impossible to distinguish the edge versus the tail at finite N.
- **(Moderately) Heavy-Tailed**, $2 < \mu < 4$: Here, the ESD $\rho_N(\lambda)$ is Heavy-Tailed / PL in the infinite limit, approaching $\rho(\lambda) \sim \lambda^{-1-\mu/2}$. In this regime, there is no bulk edge. At finite size, the global ESD can be modeled by $\rho_N(\lambda) \sim \lambda^{-(a\mu+b)}$, for all $\lambda > \lambda_{min}$, but the slope a and intercept b must be fit, as they display large finite-size effects. The maximum eigenvalues follow Frechet (not TW) statistics, with $\lambda_{max} \sim M^{4/\mu-1}(1/Q)^{1-2/\mu}$, and they have large finite-size effects. Thus, at any finite N, $\rho_N(\lambda)$ is Heavy-Tailed, but the tail decays moderately quickly.
- **(Very) Heavy-Tailed**, $0 < \mu < 2$: Here, the ESD $\rho_N(\lambda)$ is Heavy-Tailed / PL for all finite N, and as $N \to \infty$ it converges more quickly to a PL distribution with tails $\rho(\lambda) \sim \lambda^{-1-\mu/2}$. In this regime, there is no bulk edge, and the maximum eigenvalues follow Frechet (not TW) statistics. Finite-size effects exist, but they are are much smaller here than in the $2 < \mu < 4$ regime of μ.

Fitting PL distributions to ESD plots. Once we have identified PL distributions visually, we can fit the ESD to a PL in order to obtain the exponent α. We use the Clauset-Shalizi-Newman (CSN) approach (Clauset et al., 2009),

	Generative Model w/ elements from Universality class	Finite-N Global shape $\rho_N(\lambda)$	Limiting Global shape $\rho(\lambda),\ N \to \infty$	Bulk edge Local stats $\lambda \approx \lambda^+$	(far) Tail Local stats $\lambda \approx \lambda_{max}$
Basic MP	Gaussian	MP, i.e., Eqn. (1)	MP	TW	No tail.
Spiked-Covariance	Gaussian, + low-rank perturbations	MP + Gaussian spikes	MP	TW	Gaussian
Heavy tail, $4 < \mu$	(Weakly) Heavy-Tailed	MP + PL tail	MP	Heavy-Tailed*	Heavy-Tailed*
Heavy tail, $2 < \mu < 4$	(Moderately) Heavy-Tailed (or "fat tailed")	PL** $\sim \lambda^{-(a\mu+b)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet
Heavy tail, $0 < \mu < 2$	(Very) Heavy-Tailed	PL** $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet

Table 1. Basic MP theory, and the spiked and Heavy-Tailed extensions we use, including known, empirically-observed, and conjectured relations between them. Boxes marked "*" are best described as following "TW with large finite size corrections" that are likely Heavy-Tailed (Biroli et al., 2007b), leading to bulk edge statistics and far tail statistics that are indistinguishable. Boxes marked "**" are phenomenological fits, describing large ($2 < \mu < 4$) or small ($0 < \mu < 2$) finite-size corrections on $N \to \infty$ behavior. See (Davis et al., 2014; Biroli et al., 2007b;a; Péché; Auffinger et al., 2009; Edelman et al., 2016; Auffinger & Tang, 2016; Burda & Jurkiewicz, 2009; Bouchaud & Potters, 2011; Bouchaud & Mézard, 1997) for additional details.

as implemented in the python PowerLaw package (Alstott et al., 2014)

Identifying the Universality class. Given α, we identify the corresponding μ and thus which of the three Heavy-Tailed Universality classes ($0 < \mu < 2$ or $2 < \mu < 4$ or $4 < \mu$, as described in Table 1) is appropriate to describe the system. The following are particularly important points. First, observing a Heavy-Tailed ESD may indicate the presence of a scale-free DNN. This suggests that the underlying DNN is strongly-correlated, and that we need more than just a few separated spikes, plus some random-like bulk structure, to model the DNN and to understand DNN regularization. Second, this does not necessarily imply that the matrix elements of $\mathbf{W}_l$ form a Heavy-Tailed distribution. Rather, the Heavy-Tailed distribution arises since we posit it as a model of the strongly correlated, highly non-random matrix $\mathbf{W}_l$. Third, we conjecture that this is more general, and that very well-trained DNNs will exhibit Heavy-Tailed behavior in their ESD for many the weight matrices.

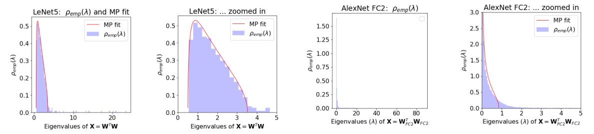
3. Empirical Results: ESDs for Existing, Pretrained DNNs

Early on, we observed that small DNNs and large DNNs have very different ESDs. For smaller models, ESDs tend to fit the MP theory well, with well-understood deviations, e.g., low-rank perturbations. For larger models, the ESDs $\rho_N(\lambda)$ almost never fit the theoretical $\rho_{mp}(\lambda)$, and they frequently have a completely different form. We use RMT to compare and contrast the ESDs of a smaller, older NN and many larger, modern DNNs. For the small model, we retrain a modern variant of one of the very early and well-known Convolutional Nets—LeNet5. For the larger, modern models, we examine selected layers from AlexNet, InceptionV3, and many other models (as distributed with pyTorch).

Example: LeNet5 (1998). LeNet5 is the prototype early model for DNNs (LeCun et al., 1998). Since LeNet5 is older, we actually recoded and retrained it. We used Keras 2.0, using 20 epochs of the AdaDelta optimizer, on the MNIST data set. This model has 100.00% training accuracy, and 99.25% test accuracy on the default MNIST split. We analyze the ESD of the FC1 Layer. The FC1 matrix $\mathbf{W}_{FC1}$ is a 2450×500 matrix, with $Q = 4.9$, and thus it yields 500 eigenvalues.

Figures 1(a) and 1(b) present the ESD for FC1 of LeNet5, with Figure 1(a) showing the full ESD and Figure 1(b) zoomed-in along the X-axis. We show (red curve) our fit to the MP distribution $\rho_{emp}(\lambda)$. Several things are striking. First, the *bulk* of the density $\rho_{emp}(\lambda)$ has a large, MP-like shape for eigenvalues $\lambda < \lambda^+ \approx 3.5$, and the MP distribution fits this part of the ESD *very* well, including the fact that the ESD just below the best fit λ^+ is concave. Second, *some eigenvalue mass is bleeding out* from the MP bulk for $\lambda \in [3.5, 5]$, although it is quite small. Third, beyond the MP bulk and this bleeding out region, are several *clear outliers, or spikes*, ranging from ≈ 5 to $\lambda_{max} \lesssim 25$. Overall, the shape of $\rho_{emp}(\lambda)$, the quality of the global bulk fit, and the statistics and crisp shape of the local bulk edge all agree well with MP theory plus a low-rank perturbation.

Example: AlexNet (2012). AlexNet was the first modern DNN (Krizhevsky et al., 2012). AlexNet resembles a scaled-up version of the LeNet5 architecture; it consists of 5 layers, 2 convolutional, followed by 3 FC layers (the last being a softmax classifier). We refer to the last 2 layers before the



(a) LeNet5, full (b) LeNet5, zoomed-in (c) AlexNet, full (d) AlexNet, zoomed-in

Figure 1. Full and zoomed-in ESD for LeNet5 (Layer FC1) and AlexNet (Layer FC2). Overlaid (in red) are fits of the MP distribution (which fit the bulk very well for LeNet5 but *not* well for AlexNet).

final softmax as layers FC1 and FC2, respectively. FC2 has a 4096×1000 matrix, with $Q = 4.096$.

Consider AlexNet FC2 (full in Figures 1(c), and zoomed-in in 1(d)). This ESD differs even more profoundly from standard MP theory. Here, we could find no good MP fit. The best MP fit (in red) does not fit the Bulk part of $\rho_{emp}(\lambda)$ well. The fit suggests there should be significantly more bulk eigenvalue mass (i.e., larger empirical variance) than actually observed. In addition, the bulk edge is indeterminate by inspection. It is only defined by the crude fit we present, and any edge statistics obviously do not exhibit TW behavior. In contrast with MP curves, which are convex near the bulk edge, the entire ESD is concave (nearly) everywhere. Here, a PL fit gives good fit $\alpha \approx 2.25$, indicating a $\mu \lesssim 3$. The shape of $\rho_{emp}(\lambda)$, the quality of the global bulk fit, and the statistics and shape of the local bulk edge are poorly-described by standard MP theory.

Empirical results for other pre-trained DNNs. We have also examined the properties of a wide range of other pre-trained models, and we have observed similar Heavy-Tailed properties to AlexNet in all of the larger, state-of-the-art DNNs, including VGG16, VGG19, ResNet50, InceptionV3, etc. Several observations can be made. First, all of our fits, except for certain layers in InceptionV3, appear to be in the range $1.5 < \alpha \lesssim 3.5$ (where the CSN method is known to perform well). Second, we also check to see whether PL is the best fit by comparing the distribution to a Truncated Power Law (TPL), as well as an exponential, stretch-exponential, and log normal distributions. In all cases, we find either a PL or TPL fits best (with a p-value ≤ 0.05), with TPL being more common for smaller values of α. Third, even when taking into account the large finite-size effects in the range $2 < \alpha < 4$, nearly all of the ESDs appear to fall into the $2 < \mu < 4$ Universality class.

Towards a Theory of Self-Regularization. For older and/or smaller models, like LeNet5, the *bulk* of their ESDs ($\rho_N(\lambda);\ \lambda \ll \lambda^+$) can be well-fit to theoretical MP density $\rho_{mp}(\lambda)$, potentially with distinct, outlying *spikes* ($\lambda > \lambda^+$). This is consistent with the Spiked-Covariance model of Johnstone (Johnstone, 2001), a simple perturbative extension of the standard MP theory. This is also reminiscent of traditional Tikhonov regularization, in that there is a "size scale" (λ^+) separating signal (spikes) from noise (bulk). This demonstrates that the DNN training process itself engineers a form of implicit *Self-Regularization* into the trained model.

For large, deep, state-of-the-art DNNs, our observations suggest that there are profound deviations from traditional RMT. These networks are reminiscent of strongly-correlated disordered-systems that exhibit Heavy-Tailed behavior. What is this regularization, and how is it related to our observations of implicit Tikhonov-like regularization on LeNet5?

To answer this, recall that similar behavior arises in strongly-correlated physical systems, where it is known that strongly-correlated systems can be *modeled* by random matrices—with entries drawn from non-Gaussian Universality classes (Sornette, 2006), e.g., PL or other Heavy-Tailed distributions. Thus, when we observe that $\rho_N(\lambda)$ has Heavy-Tailed properties, we can hypothesize that $\mathbf{W}$ is strongly-correlated,[1] and we can model it with a Heavy-Tailed distribution. Then, upon closer inspection, we find that the ESDs of large, modern DNNs behave as expected—when using the lens of Heavy-Tailed variants of RMT. Importantly, unlike the Spiked-Covariance case, which has a scale cut-off (λ^+), in these very strongly Heavy-Tailed cases, correlations appear on every size scale, and we can not find a clean separation between the MP bulk and the spikes. These observations demonstrate that modern, state-of-the-art DNNs exhibit a new form of *Heavy-Tailed Self-Regularization*.

4. 5+1 Phases of Regularized Training

In this section, we develop an operational/phenomenological theory for DNN Self-Regularization.

MP Soft Rank. We first define the *MP Soft Rank* ($\mathcal{R}_{mp}$), that is designed to capture the "size scale" of the noise part of $\mathbf{W}_l$, relative to the largest eigenvalue of $\mathbf{W}_l^T\mathbf{W}_l$. Assume that MP theory fits *at least a bulk* of $\rho_N(\lambda)$. Then, we can identify a bulk edge λ^+ and a bulk variance σ^2_{bulk}, and define the *MP Soft Rank* as the ratio of λ^+ and λ_{max}: $\mathcal{R}_{mp}(\mathbf{W}) := \lambda^+/\lambda_{max}$. Clearly, $\mathcal{R}_{mp} \in [0, 1]$; $\mathcal{R}_{mp} = 1$ for a purely random matrix; and for a matrix with an ESD with outlying spikes, $\lambda_{max} > \lambda^+$, and $\mathcal{R}_{mp} < 1$. If there is no good MP fit because the entire ESD is well-approximated by a Heavy-Tailed distribution, then we can define $\lambda^+ = 0$, in which case $\mathcal{R}_{mp} = 0$.

Visual Taxonomy. We characterize *implicit Self-Regularization*, both for DNNs during SGD training as well as for *pre-trained* DNNs, as a visual taxonomy of *5+1 Phases of Training* (RANDOM-LIKE, BLEEDING-

[1]For DNNs, these correlations arise in the weight matrices during Backprop training. That is, the weight matrices "learn" the correlations in the data.

	Operational Definition	Informal Description via Eqn. (4)	Edge/tail Fluctuation Comments	Illustration and Description
RANDOM-LIKE	ESD well-fit by MP with appropriate λ^+	$\mathbf{W}^{rand}$ random; $\|\Delta^{sig}\|$ zero or small	$\lambda_{max} \approx \lambda^+$ is sharp, with TW statistics	Fig. 2(a)
BLEEDING-OUT	ESD RANDOM-LIKE, excluding eigenmass just above λ^+	$\mathbf{W}$ has eigenmass at bulk edge as spikes "pull out"; $\|\Delta^{sig}\|$ medium	BPP transition, λ_{max} and λ^+ separate	Fig. 2(b)
BULK+SPIKES	ESD RANDOM-LIKE plus ≥ 1 spikes well above λ^+	$\mathbf{W}^{rand}$ well-separated from low-rank Δ^{sig}; $\|\Delta^{sig}\|$ larger	λ^+ is TW, λ_{max} is Gaussian	Fig. 2(c)
BULK-DECAY	ESD less RANDOM-LIKE; Heavy-Tailed eigenmass above λ^+; some spikes	Complex Δ^{sig} with correlations that don't fully enter spike	Edge above λ^+ is not concave	Fig. 2(d)
HEAVY-TAILED	ESD better-described by Heavy-Tailed RMT than Gaussian RMT	$\mathbf{W}^{rand}$ is small; Δ^{sig} is large and strongly-correlated	No good λ^+; $\lambda_{max} \gg \lambda^+$	Fig. 2(e)
RANK-COLLAPSE	ESD has large-mass spike at $\lambda = 0$	$\mathbf{W}$ very rank-deficient; over-regularization	—	Fig. 2(f)

Table 2. The 5+1 phases of learning we identified in DNN training. We observed BULK+SPIKES and HEAVY-TAILED in existing trained models (LeNet5 and AlexNet/InceptionV3, respectively; see Section 3); and we exhibited all 5+1 phases in a simple model (MiniAlexNet; see Section 6).

OUT, BULK+SPIKES, BULK-DECAY, HEAVY-TAILED, and RANK-COLLAPSE). See Table 2 for a summary. The 5+1 phases can be ordered, with each successive phase corresponding to a smaller Stable Rank / MP Soft Rank and to progressively more Self-Regularization than previous phases. Figure 2 depicts typical ESDs for each phase, with the MP fits (in red). Earlier phases of training correspond to the final state of older and/or smaller models like LeNet5 and MLP3. Later phases correspond to the final state of more modern models like AlexNet, Inception, etc. While we can describe this in terms of SGD training, this taxonomy allows us to compare different architectures and/or amounts of regularization in a trained—or even pre-trained—DNN.

Each phase is visually distinct, and each has a natural interpretation in terms of RMT. One consideration is the *global properties of the ESD*: how well all or part of the ESD is fit by an MP distriution, for some value of λ^+, or how well all or part of the ESD is fit by a Heavy-Tailed or PL distribution, for some value of a PL parameter. A second consideration is *local properties of the ESD*: the form of fluctuations, in particular around the edge λ^+ or around the largest eigenvalue λ_{max}. For example, the shape of the ESD near to and immediately above λ^+ is very different in Figure 2(a) and Figure 2(c) (where there is a crisp edge) versus Figure 2(b) (where the ESD is concave) versus Figure 2(d) (where the ESD is convex).

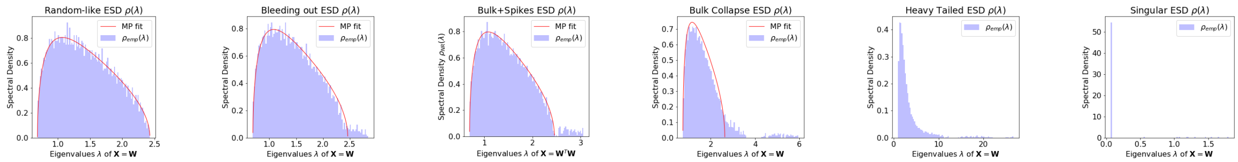
Theory of Each Phase. RMT provides more than simple visual insights, and we can use RMT to differentiate between the *5+1 Phases of Training* using simple models that qualitatively describe the shape of each ESD. We model the weight matrices $\mathbf{W}$ as "noise plus signal," where the "noise" is modeled by a random matrix $\mathbf{W}^{rand}$, with entries drawn from the Gaussian Universality class (well-described by traditional MP theory) and the "signal" is a (small or large) correction Δ^{sig}:

$$\mathbf{W} \simeq \mathbf{W}^{rand} + \Delta^{sig}. \tag{4}$$

Table 2 summarizes the theoretical model for each phase. Each model uses RMT to describe the global shape of $\rho_N(\lambda)$, the local shape of the fluctuations at the bulk edge, and the statistics and information in the outlying spikes, including possible Heavy-Tailed behaviors.

In the first phase (RANDOM-LIKE), the ESD is well-described by traditional MP theory, in which a random matrix has entries drawn from the Gaussian Universality class. In the next phases (BLEEDING-OUT, BULK+SPIKES), and/or for small networks such as LetNet5, Δ^{sig} is a relatively-small perturbative correction to $\mathbf{W}^{rand}$, and vanilla MP theory (as reviewed in Section 2.1) can be applied, as least to the bulk of the ESD. In these phases, we will *model* the $\mathbf{W}^{rand}$ matrix by a vanilla $\mathbf{W}_{mp}$ matrix (for appropriate parameters), and the MP Soft Rank is relatively large ($\mathcal{R}_{mp}(\mathbf{W}) \gg 0$). In the BULK+SPIKES phase, the model resembles a Spiked-Covariance model, and the Self-Regularization resembles Tikhonov regularization.

In later phases (BULK-DECAY, HEAVY-TAILED), and/or for modern DNNs such as AlexNet and InceptionV3, Δ^{sig} becomes more complex and increasingly dominates over



(a) RANDOM-LIKE. (b) BLEEDING-OUT. (c) BULK+SPIKES. (d) BULK-DECAY. (e) HEAVY-TAILED. (f) RANKCOLLAPSE.

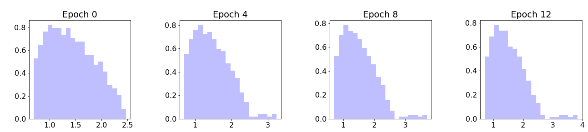
Figure 2. Taxonomy of trained models. Starting off with an initial random or RANDOM-LIKE model (2(a)), training can lead to a BULK+SPIKES model (2(c)), with data-dependent spikes on top of a random-like bulk. Depending on the network size and architecture, properties of training data, etc., additional training can lead to a HEAVY-TAILED model (2(e)), a high-quality model with long-range correlations. An intermediate BLEEDING-OUT model (2(b)), where spikes start to pull out from the bulk, and an intermediate BULK-DECAY model (2(d)), where correlations start to degrade the separation between the bulk and spikes, leading to a decay of the bulk, are also possible. In extreme cases, a severely over-regularized model (2(f)) is possible.

$\mathbf{W}^{rand}$. For these more strongly-correlated phases, $\mathbf{W}^{rand}$ is relatively much weaker, and the MP Soft Rank decreases. Vanilla MP theory is not appropriate, and instead the Self-Regularization becomes Heavy-Tailed. We will treat the noise term $\mathbf{W}^{rand}$ as small, and we will *model* the properties of Δ^{sig} with Heavy-Tailed extensions of vanilla MP theory (as reviewed in Section 2.2) to Heavy-Tailed non-Gaussian universality classes that are more appropriate to model strongly-correlated systems. In these phases, the strongly-correlated model is still regularized, but in a very non-traditional way. The final phase, the RANK-COLLAPSE phase, is a degenerate case that is a prediction of the theory.

5. Empirical Results: Detailed Analysis on Smaller Models

To validate and illustrate our theory, we analyzed MiniAlexNet,[2] a simpler version of AlexNet, similar to the smaller models used in (Zhang et al., 2016), scaled down to prevent overtraining, and trained on CIFAR10. The basic architecture consists of two 2D Convolutional layers, each with Max Pooling and Batch Normalization, giving 6 initial layers; it then has two Fully Connected (FC), or Dense, layers with ReLU activations; and it then has a final FC layer added, with 10 nodes and softmax activation. $\mathbf{W}_{FC1}$ is a 4096×384 matrix ($Q \approx 10.67$); $\mathbf{W}_{FC2}$ is a 384×192 matrix ($Q = 2$); and $\mathbf{W}_{FC3}$ is a 192×10 matrix. All models are trained using Keras 2.x, with TensorFlow as a backend. We use SGD with momentum, with a learning rate of 0.01, a momentum parameter of 0.9, and a baseline batch size of 32; and we train up to 100 epochs. We save the weight matrices at the end of every epoch, and we analyze the empirical properties of the $\mathbf{W}_{FC1}$ and $\mathbf{W}_{FC2}$ matrices.

For each layer, the matrix Entropy ($\mathcal{S}(\mathbf{W})$) gradually lowers; and the Stable Rank ($\mathcal{R}_s(\mathbf{W})$) shrinks. These decreases parallel the increase in training/test accuracies, and both metrics level off as the training/test accuracies do. These changes are seen in the ESD, e.g., see Figure 3. For layer FC1, the initial weight matrix $\mathbf{W}^0$ looks very much like an MP distribution (with $Q \approx 10.67$), consistent with a RANDOM-LIKE phase. Within a very few epochs, however, eigenvalue mass shifts to larger values, and the ESD looks like the BULK+SPIKES phase. Once the Spike(s) appear(s), substantial changes are hard to see visually, but minor changes do continue in the ESD. Most notably, λ^{max} increases from roughly 3.0 to roughly 4.0 during training, indicating further Self-Regularization, even within the BULK+SPIKES phase. Here, spike eigenvectors tend to be more localized than bulk eigenvectors. If explicit regularization (e.g., L_2 norm weight regularization or Dropout) is added, then we observe a greater decrease in the complexity metrics (Entropies and Stable Ranks), consistent with expectations, and this is casued by the eigenvalues in the spike being pulled to much larger values in the ESD. We also observe that eigenvector localization tends to be more prominent, presumably since explicit regularization can make spikes more well-separated from the bulk.



(a) Epoch 0 (b) Epoch 4 (c) Epoch 8 (d) Epoch 12

Figure 3. Baseline ESD for Layer FC1 of MiniAlexNet, during training.

6. Explaining the Generalization Gap by Exhibiting the Phases

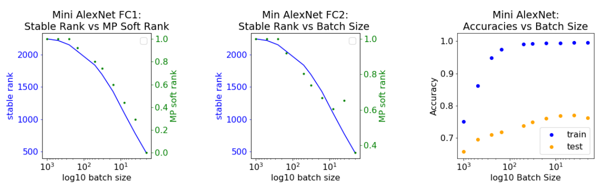
In this section, we demonstrate that we can exhibit all five of the main phases of learning by changing a single knob of the learning process. We consider the batch size since it is not traditionally considered a regularization parameter and

[2] https://github.com/deepmind/sonnet/blob/master/sonnet/python/modules/nets/alexnet.py

due to its its implications for the generalization gap.

The *Generalization Gap* refers to the peculiar phenomena that DNNs generalize significantly less well when trained with larger mini-batches (on the order of $10^3 - 10^4$) (LeCun et al., 1988; Hoffer et al., 2017; Keskar et al., 2016; Goyal et al., 2017). Practically, this is of interest since smaller batch sizes makes training large DNNs on modern GPUs much less efficient. Theoretically, this is of interest since it contradicts simplistic stochastic optimization theory for convex problems. Thus, there is interest in the question: what is the mechanism responsible for the drop in generalization in models trained with SGD methods in the large-batch regime?

To address this question, we consider here using different batch sizes in the DNN training algorithm. We trained the MiniAlexNet model, just as in Section 5, except with batch sizes ranging from moderately large to very small ($b \in \{500, 250, 100, 50, 32, 16, 8, 4, 2\}$).

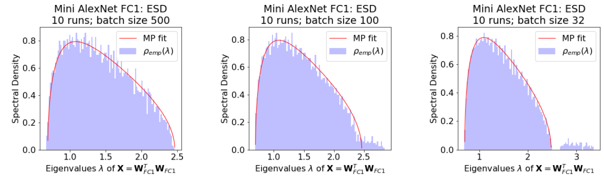


(a) Layer FC1. (b) Layer FC2. (c) Training, Test Accuracies.

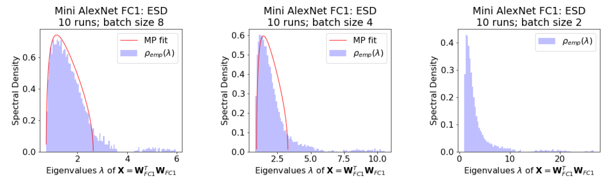
Figure 4. Varying Batch Size. Stable Rank and MP Softrank for FC1 (4(a)) and FC2 (4(b)); and Training and Test Accuracies (4(c)) versus Batch Size for MiniAlexNet.

Stable Rank, MP Soft Rank, and Training/Test Performance. Figure 4 shows the Stable Rank and MP Softrank for FC1 (4(a)) and FC2 (4(b)) as well as the Training and Test Accuracies (4(c)) as a function of Batch Size. The MP Soft Rank ($\mathcal{R}_{mp}$) and the Stable Rank ($\mathcal{R}_s$) both track each other, and both systematically *decrease* with decreasing batch size, as the test accuracy *increases*. In addition, both the training and test accuracy decrease for larger values of b: training accuracy is roughly flat until batch size $b \approx 100$, and then it begins to decrease; and test accuracy actually increases for extremely small b, and then it gradually decreases as b increases.

ESDs: Comparisons with RMT. Figure 5 shows the final ensemble ESD for each value of b for Layer FC1. We see systematic changes in the ESD as batch size b decreases. At batch size $b = 250$ (and larger), the ESD resembles a pure MP distribution with no outliers/spikes; it is RANDOM-LIKE. As b decreases, there starts to appear an outlier region. For $b = 100$, the outlier region resembles BLEEDING-OUT. For $b = 32$, these eigenvectors become well-separated from



(a) Batch Size 500. (b) Batch Size 100. (c) Batch Size 32.



(d) Batch Size 8. (e) Batch Size 4. (f) Batch Size 2.

Figure 5. Varying Batch Size. ESD for Layer FC1 of MiniAlexNet, with MP fit (in red), for an ensemble of 10 runs, for Batch Size ranging from 500 down to 2. Smaller batch size leads to more implicitly self-regularized models. We exhibit all 5 of the main phases of training by varying only the batch size.

the bulk, and the ESD resembles BULK+SPIKES. As batch size continues to decrease, the spikes grow larger and spread out more (observe the scale of the X-axis), and the ESD exhibits BULK-DECAY. Finally, at $b = 2$, extra mass from the main part of the ESD plot almost touches the spike, and the curvature of the ESD changes, consistent with HEAVY-TAILED. In addition, as b decreases, some of the extreme eigenvectors associated with eigenvalues that are not in the bulk tend to be more localized.

Implications for the generalization gap. Our results here (both that training/test accuracies decrease for larger batch sizes and that smaller batch sizes lead to more well-regularized models) demonstrate that the generalization gap phenomenon arises since, for smaller values of the batch size b, the DNN training process itself implicitly leads to stronger Self-Regularization. (This Self-Regularization can be either the more traditional Tikhonov-like regularization or the Heavy-Tailed Self-Regularization corresponding to strongly-correlated models.) That is, training with smaller batch sizes implicitly leads to more well-regularized models, and it is this regularization that leads to improved results. The obvious mechanism is that, by training with smaller batches, the DNN training process is able to "squeeze out" more and more finer-scale correlations from the data, leading to more strongly-correlated models. Large batches, involving averages over many more data points, simply fail to see this very fine-scale structure, and thus they are less able to construct strongly-correlated models characteristic of the HEAVY-TAILED phase.

References

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

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.

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. Financial applications of random matrix theory: a short review. In Akemann, G., Baik, J., and Francesco, P. D. (eds.), *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, 2011.

Bun, J., Bouchaud, J.-P., and Potters, M. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

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

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.

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.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley & Sons, 2001.

Edelman, A., Guionnet, A., and Péché, S. Beyond universality in random matrix theory. *Ann. Appl. Probab.*, 26 (3):1659–1697, 2016.

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

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

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. Technical Report Preprint: arXiv:1706.02677, 2017.

Haussler, D., Kearns, M., Seung, H. S., and Tishby, N. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2):195236, 1996.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. Technical Report Preprint: arXiv:1705.08741, 2017.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, pp. 295–327, 2001.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: generalization gap and sharp minima. Technical Report Preprint: arXiv:1609.04836, 2016.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Annual Advances in Neural Information Processing Systems 25: Proceedings of the 2012 Conference*, pp. 1097–1105, 2012.

Laloux, L., Cizeau, P., Bouchaud, J.-P., and Potters, M. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83(7):1467–1470, 1999.

Laloux, L., Cizeau, P., Potters, M., and Bouchaud, J.-P. Random matrix theory and financial correlations. *Mathematical Models and Methods in Applied Sciences*, pp. 109–11, 2005.

LeCun, Y., Bottou, L., and Orr, G. Efficient backprop in neural networks: Tricks of the trade. *Lectures Notes in Computer Science*, 1524, 1988.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Malevergne, Y. and Sornette, D. Collective origin of the coexistence of apparent RMT noise and factors in large sample correlation matrices. Technical Report Preprint: arXiv:cond-mat/0210115, 2002.

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, 2018.

Martin, C. H. and Mahoney, M. W. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. Technical Report Preprint: arXiv:1901.08278, 2019.

Péché, S. The edge of the spectrum of random matrices. Technical Report UMR 5582 CNRS-UJF, Université Joseph Fourier.

Seung, H. S., Sompolinsky, H., , and Tishby, N. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992.

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

Vapnik, V., Levin, E., and Cun, Y. L. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6 (5):851–876, 1994.

Watkin, T. L. H., Rau, A., and Biehl, M. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65(2): 499–556, 1993.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. Technical Report Preprint: arXiv:1611.03530, 2016.