ON THE POWER-LAW HESSIAN SPECTRA IN DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

It is well-known that the Hessian of deep loss landscape matters to optimization, generalization, and even robustness of deep learning. Recent works empirically discovered that the Hessian spectrum in deep learning has a two-component structure that consists of a small number of large eigenvalues and a large number of nearly-zero eigenvalues. However, the mathematical structure behind the Hessian spectra is still under-explored. To the best of our knowledge, we are the first to demonstrate that the Hessian spectra of well-trained deep neural networks exhibit simple power-law structures. Inspired by the statistical physics theories, we provide a maximum-entropy theoretical interpretation for explaining why the power-law structure exists. Our extensive experiments using the novel power-law spectral method reveal that the power-law Hessian spectra critically relate to multiple important behaviors of deep learning, including optimization, generalization, overparameterization, and overfitting.

1 INTRODUCTION

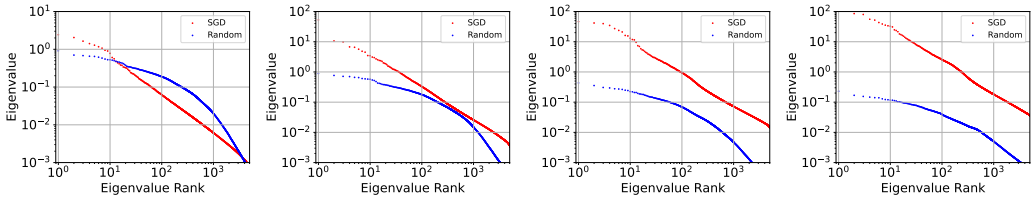
It is well-known that the Hessian matters to optimization, generalization, and even robustness of deep learning (Li et al., 2020; Ghorbani et al., 2019; Zhao et al., 2019; Jacot et al., 2019; Yao et al., 2018; Dauphin et al., 2014; Byrd et al., 2011). Deep learning usually finds flat minima that generalize well (Hochreiter & Schmidhuber, 1995; 1997). The Hessian is one of the most important measures of the minima flatness and directly relates to generalization in deep learning (Hoffer et al., 2017; Neyshabur et al., 2017; Dinh et al., 2017; Wu et al., 2017; Tsuzuku et al., 2020). Jiang et al. (2019) reported that minima-flatness-based generalization bound is still the most reliable metric in extensive experiments. Wu et al. (2017) reported that the low-complexity solutions that generalize well have a small norm of Hessian matrix with respect to model parameters. Yao et al. (2018) reported that the spectrum of the Hessian closely connects to large-batch training and adversarial robustness.

A number of works empirically studied the Hessian in deep neural networks. Some papers (Sagun et al., 2016; 2017; Wu et al., 2017) empirically reported a two-component structure that, in the context of deep learning, most eigenvalues of the Hessian are nearly zero, while a small number of eigenvalues are large. Sankar et al. (2021) revealed that the layerwise Hessian spectrum is similar to the entire Hessian spectrum. However, the theoretical mechanism behind the spectrum is under-explored.

Motivation. Why does the Hessian spectrum consist of a small number of large eigenvalues and a large number of nearly zero eigenvalues? Does an elegant mathematical structure hide behind the Hessian spectrum? Our work provides a novel approach to understanding and analyzing deep learning from a spectral perspective.

Contributions. This paper has two main contributions.

1. First, to the best of our knowledge, we are the first to empirically discover and mathematically model the power-law Hessian spectra in deep learning. We theoretically formulated a novel maximum entropy interpretation for explaining the power-law Hessian spectra.
2. Second, we propose a framework of power-law spectral analysis for deep learning. We not only reveal how the power-law spectra explain the theoretical origin of striking findings but also empirically demonstrate multiple novel insights on optimization, generalization, overparameterization, and overfitting.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10 (d) CIFAR-100

Figure 1: The power-law structure of the Hessian spectrum in deep learning. Model: LeNet. We may clearly observe that the power-law spectra generally hold for well-trained deep networks on various natural or artificial datasets, while do not hold for random neural networks. We also report that a small number of outlier eigenvalues ($\sim$10) slightly deviate from the fitted straight line.

2 The Power-Law Hessian Spectrum

In this section, we demonstrate that the Hessian spectra of well-trained deep neural networks have a simple power-law structure and how to theoretically derive the power-law structure. We also show that the discovered power-law structure provides novel insights and helps understand important properties of deep learning.

Notations. We denote the training dataset as $\{(x, y)\} = \{(x_j, y_j)\}_{j=1}^N$ drawn from the data distribution $\mathcal{S}$, the n model parameters as θ and the loss function over one data sample $\{(x_j, y_j)\}$ as $l(\theta, (x_j, y_j))$. For simplicity, we further denote the training loss as $L(\theta) = \frac{1}{N}\sum_{j=1}^N l(\theta, (x_j, y_j))$. We write the descending ordered eigenvalues of the Hessian H as $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and denote the spectral density function as $p(\lambda)$.

2.1 The Power-Law Structure and Empirical Evidence

Recent papers studied the Hessian but failed to reveal its elegant mathematical structure. To better understand the distribution of the Hessian spectrum, we first visualize the Hessian spectrum of a well-trained neural network and a randomly initialized neural network by using the Lanczos algorithm (Meurant & Strakoš, 2006; Yao et al., 2020) to estimate the eigenvalues and spectral densities. In Figure 1, we display the top 6000 eigenvalues and their corresponding rank order. Both axes are log-scale. And we surprisingly discover an approximately straight line fits the Hessian spectrum of the well-trained neural network surprisingly well, except that a small number of outliers ($\sim$10) slightly deviate from the fitted straight line. To the best of our knowledge, these fitted power-law Hessian spectra were not empirically discovered or theoretically discussed by previous papers in deep learning.

The well-fitted straight line means that the observed distribution of the Hessian eigenvalues of trained neural networks approximately obeys a power-law distribution,

$$p(\lambda) = Z_c^{-1}\lambda^{-\beta}, \tag{1}$$

where Z_c is the normalization factor. The observed eigenvalues can be considered as n samples from the power-law distribution $p(\lambda)$. We may also use a corresponding finite-sample power law for describing the observed law as

$$f_k = \frac{\lambda_k}{\mathrm{Tr}(H)} = Z_d^{-1}k^{-\frac{1}{\beta-1}}, \tag{2}$$

where f is the trace-normalized eigenvalue, k is the rank order, the trace $\mathrm{Tr}(H) = \sum_{k=1}^n \lambda_k$, and $Z_d = \sum_{k=1}^n k^{-\frac{1}{\beta-1}}$ is the normalization factor for the finite-sample power law. Note that the finite-sample power law is also called Zipf's law. This can also be approximately written as

$$\lambda_k = \lambda_1 k^{-s}, \tag{3}$$

if we let $s = \frac{1}{\beta-1}$ denote the power exponent of Zipf's law. The empirical power-law spectra suggest that the estimated $\hat{\beta}$ is close to 2 and the estimated $\hat{s}$ is close to 1.

2.2 A MAXIMUM-ENTROPY THEORETICAL INTERPRETATION

The maximum entropy principle (Guiasu & Shenitzer, 1985), also named the maximum entropy prior, states that the probability distribution which best represents the current state of knowledge about a system at equilibrium is the one with the highest entropy. This principle indicates that if we have no prior knowledge for suspecting one state over any other, then all states can be considered equally likely for a system at equilibrium.

We start our theoretical analysis with maximum entropy in deep learning. The logarithmic space volume is often regarded as a kind of entropy in statistical physics (Visser, 2013). Note that flat minima have larger space volume reflected by $\det(H^{-1})$. It means maximizing the minima space volume for better generalization may be regarded as a kind of entropy maximization principle. Following Visser (2013), we may explicitly write the volume entropy as

$$S_{\text{vol}} = \log \det(H^{-1}) = \int p(\lambda) \log \lambda d\lambda \tag{4}$$

and the spectral entropy as

$$S_{\text{p}} = -\int p(\lambda) \log p(\lambda) d\lambda, \tag{5}$$

which is the entropy of the spectral density distribution.

Considering the principle of maximum entropy with the two kinds of entropy, we need to maximize the total entropy with the spectral density normalization constraint

$$S_{\text{total}} = -\int p(\lambda) \log p(\lambda) d\lambda + \beta_{\text{vol}} \int p(\lambda) \log \lambda d\lambda - \beta_{\text{norm}} (\int p(\lambda) d\lambda - 1), \tag{6}$$

where $S_{\text{total}} = S_{\text{p}} + \beta_{\text{vol}} S_{\text{vol}}$ and β_{norm} is a Lagrange multiplier. To find the optimal distribution $p^\star(\lambda)$ that maximizes the total entropy, we require the following

$$\frac{\partial S_{\text{total}}}{\partial p(\lambda)} = -\log p(\lambda) - \beta_{\text{vol}} \log \lambda - \beta_{\text{norm}} = 0. \tag{7}$$

Thus, the optimal distribution $p^\star(\lambda)$ can be solved as

$$p^\star(\lambda) = e^{-\beta_{\text{norm}}} \lambda^{\beta_{\text{vol}}}, \tag{8}$$

which has an amazingly similar form to Equation (1) with $\beta_{\text{norm}} = \log Z_c$ and $\beta = -\beta_{\text{vol}}$.

We may interpret the power-law structure of the Hessian spectrum from two simple maximum entropy principles with the spectral density normalization constraint. It roughly means that simple rules can almost explain the power-law Hessian spectrum in deep learning. The spectra have much simpler structures than previous work expected.

Interestingly, similar well-fitted power laws have been widely discussed in neuroscience (Stringer et al., 2019) and biology (Reuveni et al., 2008; Tang & Kaneko, 2020). This is exactly our motivation to further verify and study the power-law structure of the Hessian spectrum in the context of deep learning. We verify the elegant power-law structure indeed exists in well-trained deep neural networks just like bioactive proteins. In contrast, random neural networks have no such a power-law structure, just like deactivated (denatured or unfolded) proteins. Random neural networks which have no functional ability on the given task break the power-law spectra similarly to deactivated proteins.

Statistical physics theories of neural networks (Bahri et al., 2020; Torlai & Melko, 2016; Teh et al., 2003) support that randomly initialized neural networks are far from equilibrium, while well-trained neural networks are more close to equilibrium. The equilibrium condition may distinguish well-trained neural networks and randomly initialized neural networks and explains why the power-law structure breaks without training neural networks. However, we also note that there are still arguable disputes on the equilibrium of DNNs, which is beyond the main scope of this paper.

2.3 GOODNESS-OF-FIT TEST

Following Alstott et al. (2014), we use Maximum Likelihood Estimation (MLE) (Myung, 2003) for estimating the parameter β of the fitted power-law distributions and the Kolmogorov-Smirnov Test

Table 1: The Kolmogorov-Smirnov statistics of the Hessian spectra of LeNets on various datasets. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
MNIST	LeNet	Random	0.0796	0.0430	No		
MNIST	LeNet	SGD	0.00900	0.0430	Yes	1.991 ± 0.031	1.009
Fashion-MNIST	LeNet	Random	0.0971	0.0430	No		
Fashion-MNIST	LeNet	SGD	0.0132	0.0430	Yes	1.939 ± 0.030	1.065
CIFAR-10	LeNet	Random	0.0663	0.0430	No		
CIFAR-10	LeNet	SGD	0.0279	0.0430	Yes	1.968 ± 0.031	1.033
CIFAR-100	LeNet	Random	0.0944	0.0430	No		
CIFAR-100	LeNet	SGD	0.0315	0.0430	Yes	1.908 ± 0.029	1.101

Table 2: The Kolmogorov-Smirnov statistics of the Hessian eigengaps on various datasets. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
MNIST	LeNet	SGD	0.0153	0.0430	Yes	1.550 ± 0.017	1.817
Fashion-MNIST	LeNet	SGD	0.0240	0.0430	Yes	1.520 ± 0.017	1.922

(KS Test) (Massey Jr, 1951; Goldstein et al., 2004) for statistically testing the goodness of the fit. The KS test statistic is the KS distance d_{ks} between the hypothesized (fitted) distribution and the empirical data, which measures the goodness of fitting.

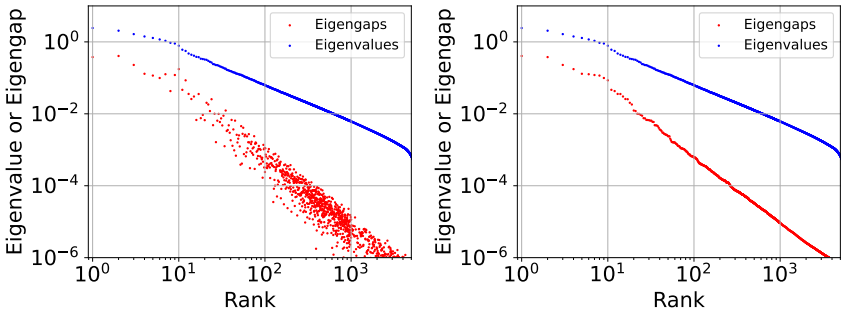
According to the practice of KS Test (Massey Jr, 1951), we first state ***the power-law hypothesis*** that the tested spectrum is power-law. If d_{ks} is higher than the critical value d_{c} at the $\alpha = 0.05$ significance level, the KS test statistically will support (not reject) the power-law hypothesis. The test results associated with Figure 1 are presented in Table 1. We leave the details and more test results (e.g., ResNet18) in Appendix A.

When we say that a spectrum is (approximately) power-law in this paper, we mean that the KS test provides positive evidence to the power-law hypothesis instead of rejecting the power-law hypothesis. Our KS test results reject the power-law hypothesis for random neural networks and do not reject the power-law hypothesis for well-trained neural networks. Moreover, when the power-law hypothesis holds, the KS distance is usually significantly smaller than the critical value d_{c}. For simplicity, the default $\alpha = 0.05$ significance level is abbreviated in the following.

Following related work on the Hessian of neural networks (Thomas et al., 2020), our empirical analysis and statistical tests mainly focused on the top (~ 1000) large eigenvalues larger than some minimal cutoff value λ_{cutoff} for three reasons. First, focusing on relatively large values is very reasonable and common in various fields' power-law studies, as real-world distributions typically follow power laws only after/large than some cutoff values (Clauset et al., 2009) for ensuring the convergence of the probability distribution. Second, researchers are usually more interested in significantly large eigenvalues which contribute more to Hessian, minima sharpness, or generalization bound (Thomas et al., 2020). Third, empirically estimating a large number of nearly zero eigenvalues is very inaccurate and expensive.

2.4 Robust and Low-Dimensional Learning Space

Deep learning happens in a low-dimensional space. Gur-Ari et al. (2018) empirically observed that deep learning (via SGD) mainly happens in a low-dimensional space during the whole training process. Ghorbani et al. (2019) studied and reported that, throughout the optimization process, large



(a) Eigenvalue Rank (b) Eigengap Rank

Figure 2: The power-law Hessian eigengaps. Model: LeNet. Datasets: MNIST. Subfigure (a) displayed the eigengaps by original rank indices sorted by eigenvalues. Subfigure (b) displayed the eigengaps by rank indices re-sorted by eigengaps. We also present the results of Fashion-MNIST in Figure 12 of Appendix E.

isolated eigenvalues rapidly appear in the spectrum, along with a surprising concentration of the gradient in the corresponding eigenspace. Xie et al. (2021a) theoretically demonstrated that the learning space is a low-dimensional subspace spanned by the eigenvectors corresponding to large eigenvalues of the Hessian, because SGD diffusion mainly happens along these principal components. Note that the low-dimensional learning space implicitly reduces deep models' complexity. However, existing work cannot explain why the low-dimensional learning space is robust during training. In this paper, robust space means that the space's dimensions are stable during training.

We try to mathematically answer this question by studying the Hessian eigengaps. We define the i-th eigengap as $\delta_k = \lambda_k - \lambda_{k+1}$. According to Equation (2), we have δ_k approximately meeting

$$\delta_k = \mathrm{Tr}(H) Z_d^{-1} (k^{-\frac{1}{\beta-1}} - (k+1)^{-\frac{1}{\beta-1}}) = \lambda_k \left[1 - (\frac{k}{k+1})^s \right]. \tag{9}$$

Interestingly, it demonstrates that eigengaps also approximately exhibit a power-law distribution when k is large. Particularly, we will have an approximate power law

$$\delta_k = \mathrm{Tr}(H) Z_d^{-1} (k+1)^{-(s+1)} \tag{10}$$

under the approximation $s \approx 1$. The power exponent $s+1$ is larger than the one in Equation (2) by 1.

Empirical analysis of the decaying Hessian eigengaps. The empirical study about the Hessian eigengaps is missing in previous papers. Our experiments filled this gap. Our experiments show that top eigengaps dominate others in deep learning similarly to eigenvalues. We further empirically verified the approximate power-law distribution of the eigengaps in Figure 2. Moreover, the observation that the power exponent of eigengaps is larger than the power exponent of eigenvalues by ~ 1 even fully matches our theoretical result by comparing Equations (2) and (10).

Note that the existence of top large eigenvalues does not necessarily indicate their gaps are also statistically large. Previous papers revealed that top eigenvalues dominate others but did not reveal if top eigengaps dominate others in deep learning. Fortunately, we theoretically and empirically demonstrate that, as rank order increases, both eigenvalues and eigengaps decay, following power-law distributions. Eigengaps even decay faster than eigenvalues due to the larger magnitude of the power exponent. We will show that this is the foundation of learning space robustness in deep learning.

Eigengaps Bound Learning Space Robustness. Based on the well-known Davis-Kahan $\sin(\Theta)$ Theorem (Davis & Kahan, 1970), we use the angle of the original eigenvector u_k and the perturbed eigenvector $\tilde{u}_k$, namely $\langle u_k, \tilde{u}_k \rangle$, to measure the robustness of space's dimensions. We directly apply Theorem 1, a useful variant of Davis-Kahan Theorem (Yu et al., 2015), to the Hessian in deep learning, which states that the eigenspace (spanned by eigenvector) robustness can be well bounded by the corresponding eigengap.

Theorem 1 (Eigengaps Bound Eigenspace Robustness (Yu et al., 2015))**.** *Suppose the true Hessian is H, the perturbed Hessian is $\tilde{H} = H + \epsilon M$, the i-th eigenvector of H is u_i , and its corresponding*

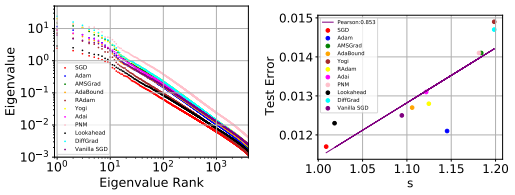


Figure 3: The power-law spectra hold across optimizers. Moreover, the slope magnitude $\hat{s}$ is an indicator of minima sharpness and a predictor of test performance. Model:LeNet. Dataset: MNIST.

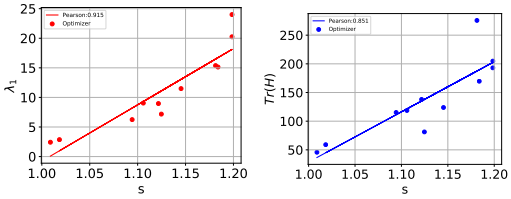


Figure 4: The slope magnitude $\hat{s}$ closely correlates to the largest Hessian eigenvalue and the Hessian trace. Model:LeNet. Dataset: MNIST.

perturbed eigenvector is $\tilde{u}_i$. Under the conditions of the Davis-Kahan Theorem, we have

$$\sin\langle u_k, \tilde{u}_k\rangle \leq \frac{2\epsilon\|M\|_{op}}{\min(\lambda_{k-1}-\lambda_k, \lambda_k-\lambda_{k+1})},$$

where $\|M\|_{op}$ is the operator norm of M.

As we have a small number of large eigengaps corresponding to the large eigenvalues, the corresponding learning space robustness has a tight upper bound. For example, given the power-law eigengaps in Equation (10), the upper bound of eigenvector robustness can be written as

$$\sup \sin\langle u_k, \tilde{u}_k\rangle = \frac{2\epsilon\|M\|_{op}(k+1)^{s+1}}{\lambda_1}, \tag{11}$$

which is relatively tight for top dimensions (small k) but becomes very loose for tailed dimensions (large k). A similar conclusion also holds given Equation (9). As $s \approx 1$ suggests, the experimental results in Figure 2 also well supports that the upper bound of $k = 1000$ is 10^4 times the upper bound of $k = 10$. This indicates that non-top eigenspace can be highly unstable during training, because δ_k can decay to nearly zero for a large k. To the best of our knowledge, we are the first to demonstrate that the robustness of low-dimensional learning space directly depends on the eigengaps of the Hessian H.

3 Related Works

A number of related works analyzed the spectral distribution of the Hessian in deep learning. Pennington & Bahri (2017) introduced an analytical framework from random matrix theory and reported that the shape of the spectrum depends strongly on the energy and the overparameterization parameter, ϕ, which measures the ratio of parameters to data points. However, Pennington & Bahri (2017) mainly evaluated single-hidden-layer networks, which limits the scope of the conclusion. A followup work (Pennington & Worah, 2018) focused on a single-hidden-layer neural network with Gaussian data and weights in the limit of infinite width. Obviously, its theoretical and empirical analysis is far from practical deep models. Jacot et al. (2019) analyzed the limiting spectrum of the Hessian in neural networks with infinite width. Fort & Scherlis (2019) analyzed the Hessian spectra of initialized neural networks. Fort & Ganguli (2019) studied the role of Hessian in learning in the low-dimensional subspace. Papyan (2019) studied the three-level hierarchical structure and outliers in Hessian spectra. Liao & Mahoney (2021) studied the Hessian spectra of more realistic nonlinear models. While a number of previous papers studied the Hessian spectrum, they failed to empirically discover the simple but important power-law structure and missed the theoretical interpretation.

4 Empirical Analysis and Discussion

In this section, we conduct extensive experimental results to explore novel behaviors of deep learning via power-law spectral analysis.

Model: LeNet (LeCun et al., 1998), Fully Connected Networks (FCN), and ResNet18 (He et al., 2016).
Dataset: MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10/100 (Krizhevsky & Hinton, 2009), and non-image Avila (De Stefano et al., 2018).

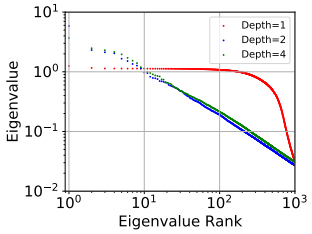


Figure 5: The power-law spectrum holds well in overparameterized deep models but disappears in the underparameterized single-layer FCN.

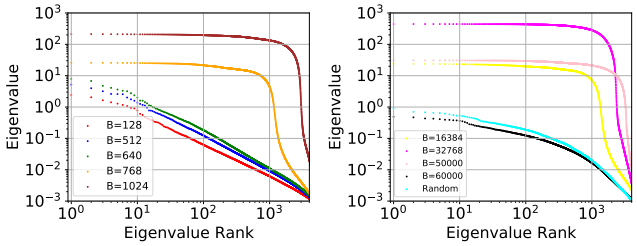


Figure 6: Batch size matters to the spectrum. We discover three phases of the Hessian spectra for large-batch training. Model:LeNet. Dataset: MNIST.

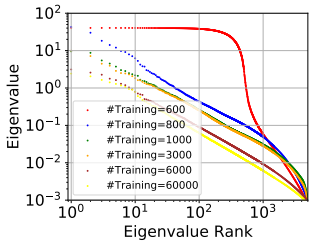


Figure 7: The spectra of LeNet on MNIST with respect to various numbers of training samples.

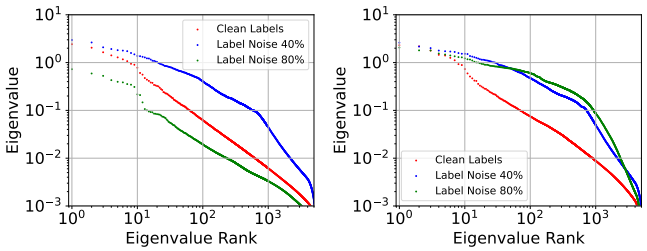


Figure 8: The spectrum in the presence of noisy labels. Left: MNIST Trainset. Right: MNIST Testset.

1. Optimization and Generalization. Figure 3 discovered that the power-law spectrum consistently holds for various popular optimizers, such as SGD, Vanilla SGD, Adam (Kingma & Ba, 2015), AMSGrad (Reddi et al., 2019), AdaBound (Luo et al., 2019), Yogi (Zaheer et al., 2018), RAdam (Liu et al., 2019), Adai (Xie et al., 2022), PNM (Xie et al., 2021b), Lookahead (Zhang et al., 2019), and DiffGrad (Dubey et al., 2019), as long as the optimizers can train neural networks well.

We find that the slope magnitude $\hat{s}$ of the fitted straight line may serve as a nice indicator of minima sharpness and a predictor of test performance. It is common to measure minima sharpness by the largest Hessian eigenvalue or the Hessian trace. A smaller $\hat{s}$ highly correlates to a smaller largest eigenvalue and a smaller trace in Figure 4. The similar observation holds on CIFAR-10 displayed in Figures 4 and 15 of Appendix E. Interestingly, we also observe that $\hat{s} \approx 1$ is common in the spectra of DNNs as well as the spectra of natural proteins.

2. Overparameterization. Figure 5 shows that the power-law spectrum holds well in overparameterized models, but disappears in underparameterized models. Overparameterization is necessary for the power-law spectrum in deep learning, while proteins are also high-dimensional. It will be interesting to study the phase transition of overparameterization and underparameterization in future.

3. Batch Size. We discover the three different phases for large-batch training via the curves in Figure 6 and the KS test results in Table 3. To our knowledge, we are the first to report the phases and sharp phase transition for large-batch training.

First, in Phase I ($B \leq 640$), moderately large-batch ($B = 512$) training indeed finds sharper minima than small-batch ($B = 128$) training, while the power-law spectrum still holds well. Power laws may guarantee that the top eigenvalues of large-batch trained networks are all larger than the corresponding eigenvalues of small-batch trained networks. The main challenge of large-batch training in Phase I is consistent with the common belief that large-batch training suffers from sharp minima and, thus, leads to bad generalization (Hoffer et al., 2017; Keskar et al., 2017). The minima sharpness measured by $\hat{s}$ obviously increases with the batch size.

Table 3: The Kolmogorov-Smirnov statistics of the Hessian spectra for various batch sizes.

Dataset	Model	Training	Batch Size	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
MNIST	LeNet	SGD	$B = 128$	0.00900	0.0430	Yes	1.991 ± 0.031	1.009
MNIST	LeNet	SGD	$B = 512$	0.00787	0.0430	Yes	1.894 ± 0.028	1.119
MNIST	LeNet	SGD	$B = 640$	0.0125	0.0430	Yes	1.838 ± 0.027	1.194
MNIST	LeNet	SGD	$B = 768$	0.278	0.0430	No		
MNIST	LeNet	SGD	$B = 1024$	0.129	0.0430	No		
MNIST	LeNet	SGD	$B = 16384$	0.249	0.0430	No		
MNIST	LeNet	SGD	$B = 32768$	0.201	0.0430	No		
MNIST	LeNet	SGD	$B = 50000$	0.139	0.0430	No		
MNIST	LeNet	SGD	$B = 60000$	0.0936	0.0430	No		

Second, in Phase II ($768 \leq B \leq 50000$), the spectrum of large-batch ($B = 1024$) trained networks does not exhibit power laws but is visually similar to the spectrum of underparameterized models in Figure 5. In Phase II, large-batch trained overparameterized models behave like underparameterized models from a spectral perspective, and, thus, can lead to bad generalization. The phase transition from Phase I to Phase II occurs in a narrow range of $640 < B < 768$, which is visually observable in Figure 6a and statistically observable in Table 3.

Third, in Phase III ($B \sim 60000$), extremely large-batch training ($B = 60000$) cannot optimize the training loss well or find the Hessian spectrum similarly to random initialized neural networks. Phase III indicates that, sometimes, bad convergence rather than sharp minima can become the main performance bottleneck in large-batch training (Xie et al., 2020), when the batch size is too large.

4. The size of training data. We evaluate the Hessian spectra over various training data sizes in Figure 7. The model trained with very limited training data finds minima with many sharp directions similarly to underparameterized models.

5. Overfitting and Noisy Labels. As DNNs overfit noisy labels easily, previous papers choose learning with noisy labels (Han et al., 2020) as an important setting for evaluating overfitting and generalization. Figure 8 shows that overfitting label noise makes the Hessian spectra less power-law on both the corrupted training dataset and the clean test dataset. In contrast, in the absence of noisy labels, the power-law spectra exist on both the training dataset and the test dataset.

6. Supplementary Results. In Appendix E, we further discussed various interesting empirical results, including modern network architecture (ResNet18), Random Labels, and Avila (a non-image dataset).

7. A Spectral Parallel of Proteins and DNNs. In this part, we make a pioneering discussion on a potential bridge between proteins and DNNs from a spectral perspective. This may connect two lines of research and draw more attention from researchers with interdisciplinary backgrounds (e.g., protein science, biophysics, etc.).

As the basic building blocks of biological intelligence, proteins work as the main executors of various vital functions, including catalysis (enzyme), transportation (carrier proteins), defense (antibodies), and so on. The polypeptide chain made up of amino acid residues can fold into its native (energy-minimum) three-dimensional structure from a random coil, which is comparable to the training of DNNs. It is worth noting that, in a long timescale, the native structure (network parameter) $\vec{r}^0$ has been gradually shaped by evolution. Due to random mutations, the native structure (energy-minimum point) and corresponding elastic network have varied. Thus, protein evolution can be recognized as an optimization process of the network parameters (Tang & Kaneko, 2021). With the second-order Taylor approximation near a minimum, for a given native structure $\vec{r}^0$, the potential energy can be calculated as:

$$E(\vec{r}^0) = E(\vec{r}^\star) + \frac{1}{2}(\vec{r}^0 - \vec{r}^\star)^\top H(\vec{r}^\star)(\vec{r}^0 - \vec{r}^\star), \tag{12}$$

in which $\vec{r}^\star$ denote the reference structure, a selected ancestral structure in the evolution, and $H(\vec{r}^\star)$ is its corresponding Hessian. In this way, the evolution of a protein becomes comparable to the training of artificial neural networks.

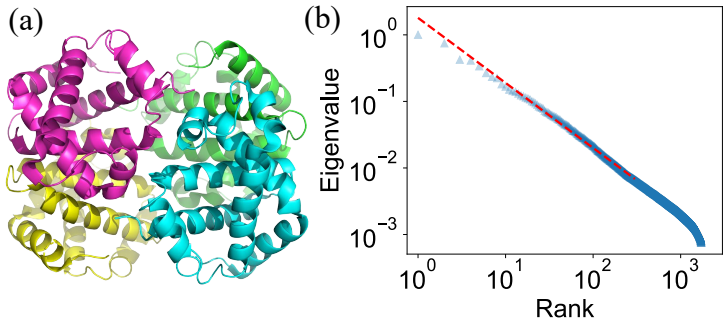


Figure 9: (a) The cartoon illustration of human deoxyhemoglobin tetramer (PDB code: 4HHB). (b) The vibrational spectrum of the protein's elastic network. The estimated power exponent $\hat{s} = 0.992$ ($\hat{\beta} = 2.008 \pm 0.032$) is very close to the estimated power exponent in deep learning.

Table 4: The estimated $\hat{s}$ for various sized proteins.

SLOPE/SIZE	$300 \leq 3N_{\rm AA} \leq 1000$	$1000 \leq 3N_{\rm AA} \leq 3000$	$3000 \leq 3N_{\rm AA} \leq 6000$
$\hat{s}$	1.050 ± 0.175	1.041 ± 0.119	1.002 ± 0.084

We use the elastic network model (ENM) of the proteins (Atilgan et al., 2001; Bahar et al., 2010) to calculate the vibrational spectra of 9166 kinds of protein molecules from the Protein Data Bank (PDB) (Berman et al., 2000). To our knowledge, this is the first large-scale spectral study via ENM for proteins. We take a human protein assembly (shown in Figure 9a) as an example to obtain the vibration spectra. The result is shown in Figure 9b.

Surprisingly, we observed $\hat{s} \approx 1$ for the protein in Figure 9 and in the spectra of total 9166 kinds of protein molecules. Each of the protein molecule has $100 \leq N_{\rm AA} \leq 2000$ amino acid residues. All of the protein spectra are statistically supported by the power-law KS tests with the estimated $mean(\hat{s}) = 1.045$ ($mean(\hat{\beta}) = 1.975$). The approximation $\hat{s} \approx 1$ also holds better for larger models (namely, proteins) in terms of the mean and the standard error, shown in Table 4. This may indicate some universal underlying mechanisms for DNNs and proteins.

A parallel maximum-entropy theory has been proposed very recently for understanding the native structure and evolution of proteins in the statistical physics community (Tang et al., 2020). Given the essential importance of protein science, our contribution to the similar power-law behaviors and theoretical interpretation may suggest a novel bridge between protein evolution and training of DNNs. In Appendix C and D, we leave more formal discussion on the spectral similarity of protein and deep learning and present more technical details.

5 CONCLUSION

In this paper, we report the power-law spectrum in deep learning. Inspired by statistical physics (Visser, 2013) and protein theory (Tang et al., 2020), we successfully formulate a novel maximum-entropy interpretation and explain why the learning space may be low dimensional and robust. The power-law spectra provide us with a powerful tool to understand and analyze deep learning. We empirically demonstrate multiple novel behaviors of deep learning, particularly deep loss landscape, beyond previous studies. Particularly, those DNNs that do not exhibit power-law decaying Hessian eigenvalues after training usually have a large number of similarly large Hessian eigenvalues and cannot generalize well. Moreover, our theoretical interpretation and large-scale empirical study on proteins suggest a likely spectral bridge to deep learning. We believe our work will inspire more theories and empirical advancements in deep learning via power-law spectral analysis in the future.

REFERENCES

Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42, 2010.

Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1), 2020.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3): 977–995, 2011.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, 27:2933–2941, 2014.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Claudio De Stefano, Marilena Maniaci, Francesco Fontanella, and A Scotto di Freca. Reliable writer identification in medieval manuscripts through page layout features: The "avila" bible case. *Engineering Applications of Artificial Intelligence*, 72:99–110, 2018.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028, 2017.

Shiv Ram Dubey, Soumendu Chakraborty, Swalpa Kumar Roy, Snehasis Mukherjee, Satish Kumar Singh, and Bidyut Baran Chaudhuri. diffgrad: An optimization method for convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 31(11):4500–4511, 2019.

Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

Stanislav Fort and Adam Scherlis. The goldilocks zone: Towards better understanding of neural network loss landscapes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3574–3581, 2019.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Michel L Goldstein, Steven A Morris, and Gary G Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2): 255–258, 2004.

Silviu Guiasu and Abe Shenitzer. The principle of maximum entropy. *The mathematical intelligencer*, 7(1):42–48, 1985.

Haiwei H Guo, Juno Choe, and Lawrence A Loeb. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences*, 101(25):9205–9210, 2004.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, pp. 529–536, 1995.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274. PMLR, 2021.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pp. 1729–1739, 2017.

Arthur Jacot, Franck Gabriel, and Clement Hongler. The asymptotic spectrum of the hessian of dnn throughout training. In *International Conference on Learning Representations*, 2019.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 190–198. SIAM, 2020.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=goEdyJ_nVQI.

Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34, 2021.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Gérard Meurant and Zdeněk Strakoš. The lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542, 2006.

Miguel A Munoz. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3):031001, 2018.

In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47 (1):90–100, 2003.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In *International Conference on Machine Learning*, pp. 5012–5021. PMLR, 2019.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pp. 2798–2806. PMLR, 2017.

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. *Advances in neural information processing systems*, 31, 2018.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *6th International Conference on Learning Representations, ICLR 2018*, 2019.

Shlomi Reuveni, Rony Granek, and Joseph Klafter. Proteins: coexistence of stability and flexibility. *Physical review letters*, 100(20):208101, 2008.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Ayaka Sakata and Kunihiko Kaneko. Dimensional reduction in evolving spin-glass model: correlation of phenotypic responses to environmental and mutational changes. *Physical Review Letters*, 124 (21):218101, 2020.

Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9481–9488, 2021.

Takuya U Sato and Kunihiko Kaneko. Evolutionary dimension reduction in phenotypic space. *Physical Review Research*, 2(1):013197, 2020.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837, 2019.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

Qian-Yuan Tang and Kunihiko Kaneko. Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis. *PLoS computational biology*, 16(2):e1007670, 2020.

Qian-Yuan Tang and Kunihiko Kaneko. Dynamics-evolution correspondence in protein structures. *Physical Review Letters*, 127(9):098103, 2021.

Qian-Yuan Tang, Yang-Yang Zhang, Jun Wang, Wei Wang, and Dante R Chialvo. Critical fluctuations in the native state of proteins. *Physical review letters*, 118(8):088102, 2017.

Qian-Yuan Tang, Tetsuhiro S Hatakeyama, and Kunihiko Kaneko. Functional sensitivity and mutational robustness of proteins. *Physical Review Research*, 2(3):033452, 2020.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.

Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3503–3513. PMLR, 2020.

Giacomo Torlai and Roger G Melko. Learning thermodynamics with boltzmann machines. *Physical Review B*, 94(16):165134, 2016.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020.

Matt Visser. Zipf's law, power laws and maximum entropy. *New Journal of Physics*, 15(4):043021, 2013.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zeke Xie, Issei Sato, and Masashi Sugiyama. Stable weight decay regularization. *arXiv preprint arXiv:2011.11152*, 2020.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021a.

Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11448–11458. PMLR, 18–24 Jul 2021b.

Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 24430–24459, 2022.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. In *Advances in Neural Information Processing Systems*, pp. 4949–4959, 2018.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590. IEEE, 2020.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Table 5: The Table of Kolmogorov-Smirnov Test Critical Values (Significance Level), which was first reported in Massey Jr (1951). If the KS distance d_{ks} is lower than a critical value, such as $\frac{1.36}{\sqrt{K}}$, we would reject the null hypothesis and accept the power-law hypothesis at the $\alpha = 0.05$ significance level. Note that K is the number of tested eigenvalues.

Sample size	$\alpha = 0.2$	$\alpha = 0.15$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
K > 35	$\frac{1.07}{\sqrt{k}}$	$\frac{1.14}{\sqrt{k}}$	$\frac{1.22}{\sqrt{k}}$	$\frac{1.36}{\sqrt{k}}$	$\frac{1.63}{\sqrt{k}}$
K = 50	0.151	0.161	0.173	0.192	0.231
K = 1000	0.0338	0.0360	0.0386	0.0430	0.0515

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in neural information processing systems*, pp. 9793–9803, 2018.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems*, 32:9597–9608, 2019.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations*, 2019.

A KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

In this section, we introduce how to conduct the Kolmogorov-Smirnov Goodness-of-Fit Test for the self-containedness purpose.

As we mentioned above, our work used Maximum Likelihood Estimation (MLE) (Myung, 2003; Clauset et al., 2009) for estimating the parameter β of the fitted power-law distributions and the Kolmogorov-Smirnov Test (KS Test) (Massey Jr, 1951; Goldstein et al., 2004) for statistically testing the goodness of the fit. The KS test statistic is the KS distance d_{ks} between the hypothesized (fitted) distribution and the empirical data, which measures the goodness of fit. Mathematically, the KS distance is defined as

$$d_{\mathrm{ks}} = \sup_{\lambda} |F^{\star}(\lambda) - \hat{F}(\lambda)|, \tag{13}$$

where $F^{\star}(\lambda)$ is the hypothesized cumulative distribution function and $\hat{F}(\lambda)$ is the empirical cumulative distribution function based on the sampled data (Goldstein et al., 2004). The estimated power exponent via MLE (Clauset et al., 2009) can be written as

$$\hat{\beta} = 1 + K \left[\sum_{i=1}^{K} \ln \left(\frac{\lambda_i}{\lambda_{\mathrm{cutoff}}} \right) \right]^{-1}, \tag{14}$$

where K is the number of tested samples and we set $\lambda_{\mathrm{cutoff}} = \lambda_k$. We note that the Powerlaw library (Alstott et al., 2014) provides a convenient tool to compute the KS distance, d_{ks}, and estimate the power exponent.

According to the practice of KS Test (Massey Jr, 1951), we first state ***the power-law hypothesis*** that the tested spectrum is power-law. If d_{ks} is higher than the critical value d_{c} at the $\alpha = 0.05$ significance level, the KS test statistically will support the power-law hypothesis (we cannot reject the power-law hypothesis). We display the critical values in Table 5.

We conducted the KS tests for all of our studied spectra. We display the KS test statistics and the estimated power exponents $\hat{\beta}$ with standard errors σ as well as the corresponding $\hat{s}$ in Tables 6, 7, 8, 9, 10,and 11. In the tables, we take the base hyperparameter setting in Appendix B as the default setting. For better visualization, we color accepting the power-law hypothesis in blue and color rejecting the power-law hypothesis (and the cause) in red.

Table 6: The Kolmogorov-Smirnov statistics of LeNet on MNIST and Fashion MNIST. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	Sample size	Setting	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
MNIST	LeNet	Random	1000	-	0.0796	0.0430	No		
MNIST	LeNet	SGD	1000	-	0.00900	0.0430	Yes	1.991 ± 0.031	1.009
MNIST	LeNet	Vanilla SGD	1000	-	0.0103	0.0430	Yes	1.914 ± 0.029	1.094
MNIST	LeNet	Adam	1000	-	0.00962	0.0430	Yes	1.873 ± 0.028	1.145
MNIST	LeNet	AMSGrad	1000	-	0.00987	0.0430	Yes	1.845 ± 0.027	1.184
MNIST	LeNet	AdaBound	1000	-	0.00889	0.0430	Yes	1.904 ± 0.029	1.106
MNIST	LeNet	Yogi	1000	-	0.00966	0.0430	Yes	1.834 ± 0.026	1.198
MNIST	LeNet	RAdam	1000	-	0.0164	0.0430	Yes	1.889 ± 0.028	1.125
MNIST	LeNet	Adai	1000	-	0.0101	0.0430	Yes	1.892 ± 0.028	1.122
MNIST	LeNet	PNM	1000	-	0.0127	0.0430	Yes	1.846 ± 0.027	1.181
MNIST	LeNet	Lookahead	1000	-	0.0101	0.0430	Yes	1.982 ± 0.031	1.018
MNIST	LeNet	DiffGrad	1000	-	0.0105	0.0430	Yes	1.834 ± 0.026	1.198
MNIST	LeNet	SGD	1000	$B = 128$	0.00900	0.0430	Yes	1.991 ± 0.031	1.009
MNIST	LeNet	SGD	1000	$B = 512$	0.00787	0.0430	Yes	1.894 ± 0.028	1.119
MNIST	LeNet	SGD	1000	$B = 640$	0.0125	0.0430	Yes	1.838 ± 0.027	1.194
MNIST	LeNet	SGD	1000	$B = 768$	0.278	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 1024$	0.129	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 8192$	0.240	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 16384$	0.249	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 32768$	0.201	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 50000$	0.139	0.0430	No		
MNIST	LeNet	SGD	1000	$B = 60000$	0.0936	0.0430	No		
MNIST	LeNet	SGD	1000	$N = 600$	0.205	0.0430	No		
MNIST	LeNet	SGD	1000	$N = 800$	0.0399	0.0430	Yes	1.995 ± 0.031	1.004
MNIST	LeNet	SGD	1000	$N = 1000$	0.0198	0.0430	Yes	2.128 ± 0.036	0.886
MNIST	LeNet	SGD	1000	$N = 3000$	0.0159	0.0430	Yes	2.091 ± 0.034	0.917
MNIST	LeNet	SGD	1000	$N = 6000$	0.0151	0.0430	Yes	2.001 ± 0.032	0.999
MNIST	LeNet	SGD	1000	40% Label Noise	0.180	0.0430	No		
MNIST	LeNet	SGD	1000	80% Label Noise	0.157	0.0430	No		
MNIST	LeNet	SGD	1000	Random Labels	0.0482	0.0430	No		
Fashion-MNIST	LeNet	Random	1000	-	0.0971	0.0430	No		
Fashion-MNIST	LeNet	SGD	1000	-	0.0132	0.0430	Yes	1.939 ± 0.030	1.065
MNIST	LeNet	SGD	1000	Eigengap	0.0153	0.0430	Yes	1.550 ± 0.017	1.817
Fashion-MNIST	LeNet	SGD	1000	Eigengap	0.0240	0.0430	Yes	1.520 ± 0.017	1.922

Table 7: The Kolmogorov-Smirnov statistics of LeNet on CIFAR-10 and CIFAR-100. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	Sample size	Setting	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
CIFAR-10	LeNet	Random	1000	-	0.0663	0.0430	No		
CIFAR-10	LeNet	SGD	1000	-	0.0279	0.0430	Yes	1.968 ± 0.031	1.033
CIFAR-10	LeNet	Vanilla SGD	1000	-	0.0276	0.0430	Yes	1.935 ± 0.030	1.069
CIFAR-10	LeNet	Adam	1000	-	0.0269	0.0430	Yes	1.806 ± 0.025	1.241
CIFAR-10	LeNet	AMSGrad	1000	-	0.0232	0.0430	Yes	1.786 ± 0.025	1.271
CIFAR-10	LeNet	AdaBound	1000	-	0.0297	0.0430	Yes	1.901 ± 0.028	1.110
CIFAR-10	LeNet	Yogi	1000	-	0.0184	0.0430	Yes	1.806 ± 0.025	1.241
CIFAR-10	LeNet	RAdam	1000	-	0.0163	0.0430	Yes	1.733 ± 0.023	1.363
CIFAR-10	LeNet	Adai	1000	-	0.0310	0.0430	Yes	1.918 ± 0.029	1.090
CIFAR-10	LeNet	PNM	1000	-	0.0347	0.0430	Yes	1.911 ± 0.029	1.098
CIFAR-10	LeNet	Lookahead	1000	-	0.0358	0.0430	Yes	1.964 ± 0.030	1.037
CIFAR-10	LeNet	DiffGrad	1000	-	0.0303	0.0430	Yes	1.803 ± 0.024	1.236
CIFAR-100	LeNet	Random	1000	-	0.0944	0.0430	No		
CIFAR-100	LeNet	SGD	1000	-	0.0315	0.0430	Yes	1.908 ± 0.029	1.101
CIFAR-100	LeNet	Vanilla SGD	1000	-	0.0379	0.0430	Yes	1.903 ± 0.029	1.108
CIFAR-100	LeNet	SGD	1000	Evaluated on CIFAR-10	0.0306	0.0430	Yes	1.913 ± 0.029	1.095

B EXPERIMENTAL SETTINGS

Computational environment. The experiments are conducted on a computing cluster with NVIDIA® V100 GPUs and Intel® Xeon® CPUs.

Table 8: The Kolmogorov-Smirnov statistics of FCN. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	Sample size	Setting	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
Avila	2Layer-FCN	SGD	50	-	0.0683	0.176	Yes	1.604 ± 0.085	1.656
MNIST	1Layer-FCN	Random	1000	-	0.185	0.0430	No		
MNIST	1Layer-FCN	SGD	1000	-	0.241	0.0430	No		
MNIST	2Layer-FCN	Random	1000	-	0.129	0.0430	No		
MNIST	2Layer-FCN	SGD	1000	-	0.0112	0.0430	Yes	2.209 ± 0.038	0.827
MNIST	4Layer-FCN	Random	1000	-	0.0628	0.0430	No		
MNIST	4Layer-FCN	SGD	1000	-	0.0141	0.0430	Yes	2.201 ± 0.038	0.833
MNIST	2Layer-FCN	SGD	1000	Width=10	0.149	0.0430	No		
MNIST	2Layer-FCN	SGD	1000	Width=20	0.185	0.0430	No		
MNIST	2Layer-FCN	SGD	1000	Width=30	0.0656	0.0430	No		
MNIST	2Layer-FCN	SGD	1000	Width=50	0.0187	0.0430	Yes	2.138 ± 0.028	0.879
MNIST	2Layer-FCN	SGD	1000	Width=70	0.0376	0.0430	Yes	2.271 ± 0.030	0.787
MNIST	2Layer-FCN	SGD	1000	Width=100	0.0112	0.0430	Yes	2.209 ± 0.038	0.827

Table 9: The Kolmogorov-Smirnov statistics of ResNet18. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	Sample size	Setting	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
CIFAR-10	ResNet18	Random	50	-	0.334	0.176	No		
CIFAR-10	ResNet18	SGD	50	-	0.0803	0.176	Yes	2.146 ± 0.162	0.873
CIFAR-10	ResNet18	Vanilla SGD	50	-	0.0891	0.176	Yes	2.193 ± 0.169	0.838
CIFAR-10	ResNet18	Adam	50	-	0.0478	0.176	Yes	2.062 ± 0.149	0.950
CIFAR-10	ResNet18	AMSGrad	50	-	0.0542	0.176	Yes	2.041 ± 0.147	0.961
CIFAR-10	ResNet18	AdaBound	50	-	0.0588	0.176	Yes	2.029 ± 0.146	0.971
CIFAR-10	ResNet18	Yogi	50	-	0.116	0.176	Yes	1.915 ± 0.129	1.092
CIFAR-10	ResNet18	RAdam	50	-	0.168	0.176	Yes	1.794 ± 0.1112	1.259
CIFAR-10	ResNet18	Adai	50	-	0.103	0.176	Yes	2.183 ± 0.167	0.845
CIFAR-10	ResNet18	PNM	50	-	0.138	0.176	Yes	2.132 ± 0.160	0.884
CIFAR-10	ResNet18	Lookahead	50	-	0.110	0.176	Yes	2.098 ± 0.155	0.911
CIFAR-10	ResNet18	DiffGrad	50	-	0.068	0.176	Yes	2.055 ± 0.149	0.948
CIFAR-10	ResNet18	SGD	50	$B = 512$	0.0561	0.176	Yes	2.146 ± 0.151	0.936
CIFAR-10	ResNet18	SGD	50	$B = 1024$	0.0647	0.176	Yes	2.076 ± 0.152	0.929
CIFAR-10	ResNet18	SGD	50	$B = 1152$	0.0598	0.176	Yes	2.060 ± 0.150	0.944
CIFAR-10	ResNet18	SGD	50	$B = 1280$	0.331	0.176	No		
CIFAR-10	ResNet18	SGD	50	$B = 2048$	0.334	0.176	No		
CIFAR-10	ResNet18	SGD	50	$B = 4096$	0.334	0.176	No		
CIFAR-10	ResNet18	SGD	50	$B = 16384$	0.343	0.176	No		
CIFAR-100	ResNet18	Random	50	-	0.373	0.176	No		
CIFAR-100	ResNet18	SGD	50	-	0.108	0.176	Yes	2.299 ± 0.184	0.770

B.1 Models, Datasets, and Optimizers

Models: LeNet (LeCun et al., 1998), Fully Connected Networks (FCN), and ResNet18 (He et al., 2016). Particularly, we used one-layer FCN, two-layer FCN, four-layer FCN, which have 100 neurons for each hidden layer and use ReLu activations.

Datasets: MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10/100 (Krizhevsky & Hinton, 2009), and non-image Avila (De Stefano et al., 2018).

Optimizers: SGD, Vanilla SGD, Adam (Kingma & Ba, 2015), AMSGrad (Reddi et al., 2019), AdaBound (Luo et al., 2019), Yogi (Zaheer et al., 2018), RAdam (Liu et al., 2019), Adai (Xie et al., 2022), PNM (Xie et al., 2021b), Lookahead (Zhang et al., 2019), and DiffGrad (Dubey et al., 2019).

B.2 Image classification on MNIST and Fashion-MNIST

Data Preprocessing For MNIST and Fashion-MNIST: We perform the common per-pixel zero-mean unit-variance normalization.

Table 10: The Kolmogorov-Smirnov statistics of LNN. The estimated power exponent $\hat{\beta}$ and slope magnitude $\hat{s}$ are also displayed.

Dataset	Model	Training	Sample size	Setting	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
MNIST	4Layer-LNN	SGD	1000	-	0.122	0.0430	No		
MNIST	4Layer-LNN	SGD	1000	w/ BatchNorm	0.0411	0.0430	Yes	2.190 ± 0.037	0.840
MNIST	4Layer-LNN	SGD	1000	w/ ReLu	0.0127	0.0430	Yes	2.054 ± 0.033	0.949

Table 11: The Kolmogorov-Smirnov Test Statistics of Protein.

Protein	Sample size	d_{ks}	d_{c}	Power-Law	$\hat{\beta} \pm \sigma$	$\hat{s}$
4HHB	1000	0.0145	0.0430	Yes	2.008 ± 0.032	0.992

Hyperparameter Settings: We select the optimal learning rate for each experiment from $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ for SGD and use the default learning rate for adaptive gradient methods. In the experiments on MNIST and Fashion-MNIST: $\eta = 0.1$ for SGD, Vanilla SGD, Adai, PNM, and Lookahead; $\eta = 0.1$ for Vanilla SGD;$\eta = 0.001$ for Adam, AMSGrad, AdaBound, Yogi, RAdam, and DiffGrad.

We train neural networks for 50 epochs on MNIST and 200 epochs on Fashion-MNIST. For the learning rate schedule, the learning rate is divided by 10 at the epoch of 40% and 80%. The batch size is set to 128 for MNIST and Fashion-MNIST, unless we specify it otherwise.

The strength of weight decay defaults to $\lambda = 0.0005$ as the baseline for all optimizers unless we specify it otherwise.

We set the momentum hyperparameter $\beta_1 = 0.9$ for SGD and adaptive gradient methods which involve in Momentum. As for other optimizer hyperparameters, we apply the default settings directly.

B.3 Image classification on CIFAR-10 and CIFAR-100

Data Preprocessing For CIFAR-10 and CIFAR-100: We perform the common per-pixel zero-mean unit-variance normalization, horizontal random flip, and 32×32 random crops after padding with 4 pixels on each side.

Hyperparameter Settings: We select the optimal learning rate for each experiment from $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ for SGD and use the default learning rate for adaptive gradient methods. In the experiments on CIFAF-10 and CIFAR-100: $\eta = 1$ for Vanilla SGD, Adai, and PNM; $\eta = 0.1$ for SGD (with Momentum) and Lookahead; $\eta = 0.001$ for Adam, AMSGrad, AdaBound, Yogi, RAdam, and DiffGrad. For the learning rate schedule, the learning rate is divided by 10 at the epoch of $\{80, 160\}$ for CIFAR-10 and $\{100, 150\}$ for CIFAR-100, respectively. The batch size is set to 128 for both CIFAR-10 and CIFAR-100, unless we specify it otherwise.

The strength of weight decay is default to $\lambda = 0.0005$ as the baseline for all optimizers unless we specify it otherwise. Recent work Xie et al. (2020) found that popular optimizers with $\lambda = 0.0005$ often yields test results than $\lambda = 0.0001$ on CIFAR-10 and CIFAR-100.

We set the momentum hyperparameter $\beta_1 = 0.9$ for SGD with Momentum. As for other optimizer hyperparameters, we apply the default hyperparameter settings directly.

B.4 Learning with noisy labels

We trained LeNet via SGD (with Momentum) on corrupted MNIST with various (asymmetric) label noise. We followed the setting of Han et al. (2018) for generating noisy labels for MNIST. The

symmetric label noise is generated by flipping every label to other labels with uniform flip rates $\{40\%, 80\%\}$. In this paper, when we talk about label noise, we mean symmetric label noise.

We also randomly shuffle the labels of MNIST to produce MNIST with random labels, which has little knowledge behind the pairs of instances and labels.

C A BRIDGE TO PROTEIN SCIENCE

In this section, we discuss recent advancements in protein science and how it inspired our discussions on deep learning.

As the basic building blocks of biological intelligence, proteins work as the main executors of various vital functions, including catalysis (enzyme), transportation (carrier proteins), defense (antibodies), and so on. The polypeptide chain made up of amino acid residues can fold into its native (energy-minimum) three-dimensional structure from a random coil. For most proteins, their "correct" native structures are essential to their functions. These structures (denote as $\vec{r}^0$) are not static. In contrast, they can perform their intrinsic dynamics due to the perturbations from the milieu. When external thermal noise act as force $\vec{\xi}$, the protein's deformation $\Delta\vec{r} = \vec{r} - \vec{r}^0$ can be estimated as: $\Delta\vec{r} = H^{-1}\vec{\xi}$, where H is the elasticity matrix, i.e., the Hessian of the potential energy landscape. Such a framework is known as the elastic network model of the proteins (Atilgan et al., 2001; Bahar et al., 2010). When take the external thermal noises $\vec{\xi}$ as the *input variable*, the deformation $\Delta\vec{r}$ can be recognized as the *output* of a trained network, and the native structure $\vec{r}^0$ encoding the elastic network corresponds to the *parameters* of the network.

Proteins as accurate and robust learners. It is worth noting that, in a long timescale, the native structure (network parameter) $\vec{r}^0$ has been gradually shaped by evolution. Due to random mutations, the native structure (energy-minimum point) and corresponding elastic network have varied. Thus, protein evolution can be recognized as an optimization process of the network parameters (Tang & Kaneko, 2021). With the second-order Taylor approximation near a minimum, for a given native structure $\vec{r}^0$, the potential energy can be calculated as:

$$E(\vec{r}^0) = E(\vec{r}^\star) + \frac{1}{2}(\vec{r}^0 - \vec{r}^\star)^\top H(\vec{r}^\star)(\vec{r}^0 - \vec{r}^\star), \tag{15}$$

in which $\vec{r}^\star$ denote the reference structure, a selected ancestral structure in the evolution, and $H(\vec{r}^\star)$ is its corresponding Hessian. In this way, the evolution of a protein becomes comparable to the training of artificial neural networks. The potential energy E corresponds to the loss function L, while native structure $\vec{r}^0$ corresponds to the model parameters θ.

(1) Accuracy. Proteins can respond with high susceptibility when perturbed by the environment, and some can undergo significant structural changes (Tang et al., 2017). The noise-induced motions are highly anisotropic, with amino acid residues moving collectively in specific directions. These movements are usually related to the protein functions (Bahar et al., 2010). It is analogous to the model's generalization ability which describes how a model (protein) can deal with new data (different external noises) and make accurate predictions (specific functional dynamics).

(2) Robustness. In protein evolution, it is the functions of proteins that act as constraints, so essential functions must withstand most mutations (Guo et al., 2004; Tang & Kaneko, 2021). It is observed that, when the environment is stable, organisms tend to evolve in a convergent direction (Sato & Kaneko, 2020; Sakata & Kaneko, 2020). Upon recognizing the protein as a learner, it is remarkable that the gradients (direction of evolution) on the loss landscape remain relatively stable. During evolution, most mutations do not affect the direction of the principal-component vectors (or the low-dimensional subspaces spanned by these vectors) related to the functions of the protein. This idea aligns with the discussions on the eigengaps in previous sections.

Power law and criticality. We take a protein assembly (shown in Figure 9a) as an example to conduct normal mode analysis and obtain the vibration spectrum. This calculation is based on the elastic network model (See Appendix D). The result is shown in Figure 9b. We evaluate 9166 kinds of proteins in Section 4. While recent studies (Reuveni et al., 2008; Tang & Kaneko, 2020; Tang et al., 2020) implicitly or explicitly suggested that the vibration spectrum of proteins exhibits a power-law distribution, we are the first to conduct large-scale statistical tests on the power-law spectra for protein. The power-law behaviors suggest the parallels between protein evolution and deep learning.

Interestingly, similar power laws were observed in various kinds of other natural systems, including brains, bird flocks, insect swarms, and so on (Munoz, 2018). In statistical physics, power laws act as the hallmark of "critical point" between ordered (robustness) and disordered (plasticity) states.

A Large-Scale Study on 9166 kinds of Proteins. Surprisingly, we observed $\hat{s} \approx 1$ for the protein in Figure 9 and in the spectra of total 9166 kinds of protein molecules from the Protein Data Bank (PDB) (Berman et al., 2000). Each of the protein molecule has $100 \leq N_{\mathrm{AA}} \leq 2000$ amino acid residues. All of the protein spectra successfully passed the power-law KS tests with $mean(\hat{s}) = 1.045$ ($mean(\hat{\beta}) = 1.975$). The approximation $\hat{s} \approx 1$ holds even better for large proteins in terms of the mean and the standard error, shown in Table 4. We conjecture there may exist some universal underlying mechanisms for deep learning and protein.

D The Spectral Analysis of Proteins

The elastic network models are widely applied to predict and characterize the slow global dynamics of a wide range of proteins and bio-machineries (Bahar et al., 2010; Tang & Kaneko, 2020). By describing the proteins as mass-and-spring networks, the elastic network models can capture the functional motions of proteins with minimal computational resources by focusing on the movement of proteins nearby the native structure. The movements of the proteins are described as the linear vibrations around the energy minimum of the energy landscape. It is worth noting that the model is not applicable for the dynamics far from the energy minimum, such as protein folding and unfolding problems.

D.1 Anisotropic Network Model (ANM)

In this work, we employ a typical form of the ENM, the Anisotropic Network Model (ANM), to calculate the vibrational spectrum of proteins (Atilgan et al., 2001; Bahar et al., 2010). Previous research has shown that not only can ANM accurately reproduce the movements of residues as determined by experiments, but the model also fits well with the data on protein structure evolution (Tang & Kaneko, 2021).

To introduce the model settings of ANM, let us first consider the sub-system consisting of two nodes (amino acid residues) i and j connected with a harmonic spring. The coordinates of the two nodes are $\vec{r}_i = [x_i, y_i, z_i]$ and $\vec{r}_j = [x_j, y_j, z_j]$, respectively; and their native-state coordinates are $\vec{r}_i^{\,0} = [x_i^0, y_i^0, z_i^0]$ and $\vec{r}_j^{\,0} = [x_j^0, y_j^0, z_j^0]$. When the equilibrium distance between them is given by $s_{ij}^0 = |\vec{r}_i^{\,0} - \vec{r}_j^{\,0}| = \sqrt{(x_i^0 - x_j^0)^2 + (y_i^0 - y_j^0)^2 + (z_i^0 - z_j^0)^2}$, and the instantaneous distance is given by $s_{ij} = |\vec{r}_i - \vec{r}_j| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$, then the potential of such a sub-system can be given as:

$$V_{ij} = \frac{\kappa}{2}\left(s_{ij} - s_{ij}^0\right)^2, \tag{16}$$

where κ denotes the spring constant. Then, around the energy minimum, the second-order derivative of the V_{ij} with respect to $\vec{r}_i$ can be given as:

$$\frac{\partial^2 V_{ij}}{\partial x_i^2} = \frac{\partial^2 V_{ij}}{\partial x_j^2} = \frac{\kappa}{s_{ij}^2}(x_j - x_i)^2, \tag{17}$$

$$\frac{\partial^2 V_{ij}}{\partial x_i \partial y_j} = \frac{\kappa}{s_{ij}^2}(x_j - x_i)(y_j - y_i). \tag{18}$$

Therefore, the corresponding Hessian matrix entries H_{ij} can be given as:

$$H_{ij} = -\frac{\kappa}{s_{ij}^2}\begin{bmatrix} x_j - x_i \\ y_j - y_i \\ z_j - z_i \end{bmatrix}[x_j - x_i, y_j - y_i, z_j - z_i]. \tag{19}$$

In this way, the $3N_{\mathrm{AA}} \times 3N_{\mathrm{AA}}$ Hessian H can be recognized as the direct sum of the 3×3 Hessian matrix H_{ij} and $N_{\mathrm{AA}} \times N_{\mathrm{AA}}$ elasticity matrix Γ. That is, the $3N_{\mathrm{AA}} \times 3N_{\mathrm{AA}}$ Hessian matrix H can

be recognized as an $N_{\rm AA} \times N_{\rm AA}$ matrix with entries of 3×3 matrices

$$H_{ij} = -\frac{\kappa \cdot \Gamma_{ij}}{s_{ij}^2} \begin{bmatrix} x_j - x_i \\ y_j - y_i \\ z_j - z_i \end{bmatrix} [x_j - x_i, y_j - y_i, z_j - z_i]. \tag{20}$$

Here, matrix Γ is defined according to the residue-residue contact topology of the native structure. In ANM, only spatial-neighboring residues are considered to be connected. For a pair of amino acid residues (i and j), if their mutual distance $r_{ij} \leq r_{\rm C}$, then $\Gamma_{ij} = -1$; if $r_{ij} > r_{\rm C}$, then $\Gamma_{ij} = 0$; and for the diagonal elements, $\Gamma_{ii} = -\sum_{j \neq i} \Gamma_{ij} = -k_i$, where k_i denote the degree of node i. Note that in a graph theory perspective, matrix Γ is also known as the graph Laplacian (or the Kirchhoff matrix) of the residue-residue contact network. In ANM with homogenous contact strength ($\kappa = 1$), the only control parameter is the cutoff distance $r_{\rm C}$. In the calculation, we take $r_{\rm C} = 9.0$ Å.

For a protein consisting of $N_{\rm AA}$ amino acid residues, the total degrees of freedom is $3N_{\rm AA}$. Among them, there are 6 degrees of freedom related to the rigid body motion, say, three-dimensional translational motions and three-dimensional rotation. Therefore, to fully describe the structural fluctuations of a protein, one need in total $n = 3N_{\rm AA} - 6$ parameters.

D.2 From ANM to SGD: Thermal noise vs. structured noise

Although the energy landscape around a protein's native state is similar to the loss landscape of a deep neural network, the corresponding noises are entirely different. Protein dynamics are driven by thermal noises, and deep learning is driven by structured noises. Due to these two types of noise, the Hessian matrix and the covariance matrix have different relationships.

For a protein molecule driven by the thermal noises, according to the Boltzmann distribution, the probability of a structure $\Delta\vec{r}$ is given as:

$$p(\Delta\vec{r}) \sim e^{-\frac{1}{2}\sum_{i,j=1}^{N_{\rm AA}} \Delta\vec{r}_i \cdot H_{ij} \cdot \Delta\vec{r}_j}, \tag{21}$$

in which $\Delta\vec{r}_i = \vec{r}_i - \vec{r}_i^0$ and $\Delta\vec{r}_j = \vec{r}_j - \vec{r}_j^0$ are three-dimensional vectors describing the displacements of residues i and j. It is worth noting that $p(\Delta\vec{r})$ is a multivariate Gaussian distribution with covariance matrix $C \sim H^{-1}$, where $C_{ij} = \langle \Delta\vec{r}_i \cdot \Delta\vec{r}_j \rangle$.

However, as discussed in the main text, for SGD, the covariance matrix is proportional to the Hessian matrix: $C_{\rm sgd}(\theta) \sim H(\theta)$. Due to this difference, it is the spectrum of the inverse (or pseudoinverse) of Hessian H^{-1} that should be applied to compare with the Hessian matrices in deep learning.

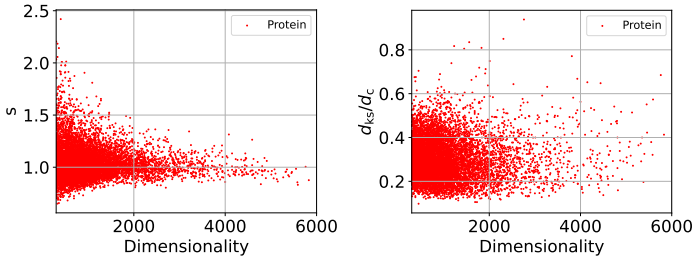
D.3 Similarities in Hessian spectra

By diagonalizing the Hessian matrix H, we can obtain all the nonzero eigenvalues and the corresponding eigenvectors describing the motions of the protein, i. e., $H = V\tilde{\Lambda}V^\top$, in which the eigenvalues $\tilde{\Lambda} = {\rm diag}[\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots \tilde{\lambda}_n]$ ($\tilde{\lambda}_1 \leq \tilde{\lambda}_2, \leq \cdots \leq \tilde{\lambda}_n$) and eigenvectors $V = [v_1, v_2, \cdots v_n]^\top$.

Then, for the inverse Hessian H^{-1}, its nonzero eigenvalues $\sigma_i = 1/\tilde{\lambda}_i$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. In Fig. 9b, as an example, the rank-size distribution of σ_i vs i is plotted. Here, we normalize all the eigenvalues as $\hat{\sigma}_i = \sigma_i/\sigma_1$. The estimated power exponent (slope magnitude) $\hat{s} = 0.992$ (corresponding to $\hat{\beta} = 2.008 \pm 0.032$). This result clearly demonstrates the similarity between the vibrational spectrum of proteins and the Hessian spectrum in deep learning.

D.4 Protein Dataset

In this work, we evaluated the vibrational spectra of 9166 kinds of protein molecules from the Protein Data Bank (PDB) (Berman et al., 2000). The studied proteins have high-resolution and clean structures, are large enough, and have enough diversity. More precisely, the structures of these proteins were all determined via high-resolution X-ray diffraction (≤ 2.0Å) without DNA, RNA, or hybrid structures involved. Their chain lengths (number of amino acid residues) are $100 \leq N_{\rm AA} \leq 2000$. Their spectra have no zero eigenvalues except for the six modes corresponding to the translational and rotational motions. Every two proteins share less than 30% sequence similarity. We choose the $\lambda_{\rm cutoff}$ as the top $\frac{1}{10}$ largest eigenvalue in each vibrational spectrum for corresponding proteins.



(a) The Slope Magnitude (b) Goodness-of-fit

Figure 10: The KS statistics of the spectra of 9166 kinds of protein molecules.

Table 12: The validity of the approximation $s \approx 1$ positively correlates with the sizes (chain lengths) of proteins.

Estimated Parameter	$300 \leq 3N_{\text{AA}} \leq 1000$	$1000 \leq 3N_{\text{AA}} \leq 3000$	$3000 \leq 3N_{\text{AA}} \leq 6000$	$300 \leq 3N_{\text{AA}} \leq 6000$
$\hat{\beta}$	1.976 ± 0.143	1.973 ± 0.103	2.004 ± 0.080	1.975 ± 0.128
$\hat{s}$	1.050 ± 0.175	1.041 ± 0.119	1.002 ± 0.084	1.045 ± 0.155
$d_{\text{ks}}/d_{\text{c}}$	0.304	0.292	0.317	0.300

We present the KS statistics of the protein spectra in Figure 10. Note that, if $\frac{d_{\text{ks}}}{d_{\text{c}}} < 1$, we would reject the null hypothesis and accept the power-law hypothesis. All of the protein spectra successfully passed the power-law KS tests with $mean(\hat{s}) = 1.045$ ($mean(\hat{\beta}) = 1.975 \pm 0.128$). We also notice that $\hat{s} \approx 1$ holds even better for larger proteins, which is supported by Table 12.

E Supplementary Empirical Results

We compared the spectrum computed via Power Iteration Algorithm and Lanczos Algorithm in Figure 11. It shows the spectrum via Power Iteration Algorithm is highly consistent with the spectrum via Lanczos Algorithm. It also demonstrates that the power-law spectrum is caused by the properties of deep learning rather than the stochasticity of Lanczos Algorithm.

We presented the power-law eigengaps on Fashion-MNIST in Figure 12. It shows that the power-law eigengaps on Fashion-MNIST are highly consistent with the power-law eigengaps on MNIST.

We presented the power-law spectrum of the covariance matrix of stochastic gradient noise of FCN on MNIST in Figure 19. As the inverses of the power-law variables are power-law, the covariance

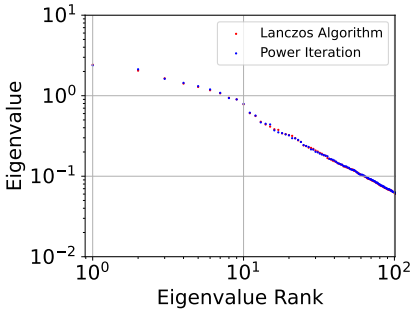
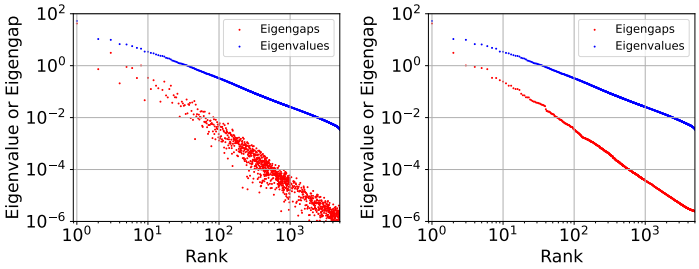


Figure 11: The spectrum via Power Iteration Algorithm is highly consistent with the spectrum via Lanczos Algorithm. It also shows that the power-law spectrum is caused by the properties of deep learning rather than the stochasticity of Lanczos Algorithm.



(a) Eigenvalue Rank (b) Eigengap Rank

Figure 12: The power-law Hessian eigengaps in deep learning. Model: LeNet. Datset: Fashion-MNIST. Subfigure (a) displayed the eigengaps by original rank indices sorted by eigenvalues. Subfigure (b) displayed the eigengaps by rank indices re-sorted by eigengaps.

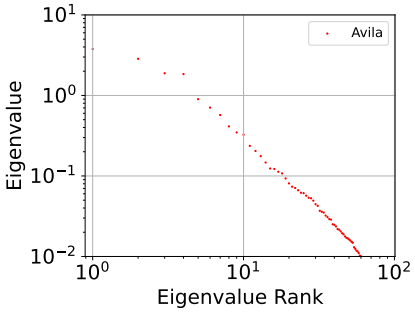


Figure 13: The spectrum of FCN on Avila Dataset. It shows that the power-law spectrum of neural networks may also exist in non-image datasets.

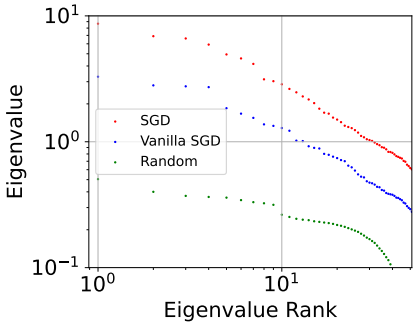


Figure 14: The power-law spectra of ResNet18 on CIFAR-10. It shows that the power-law spectrum of neural networks may also exist in modern neural network architectures (ResNet) as well as simple CNNs/FCNs.

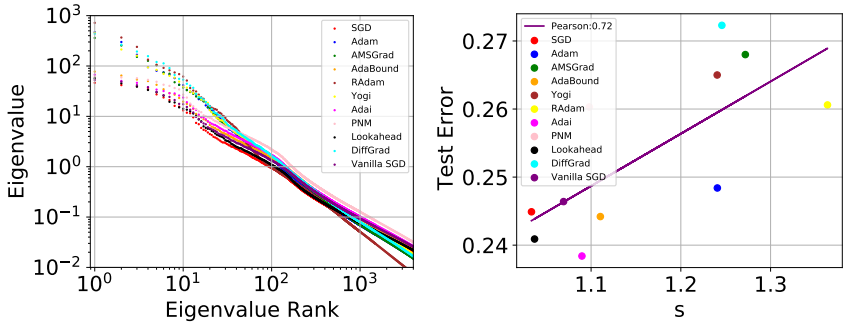


Figure 15: The power-law spectra hold across optimizers. Moreover, the slope magnitude $\hat{s}$ is an indicator of minima sharpness and a predictor of test performance. Model:LeNet. Dataset: CIFAR-10.

spectrum shows heavy-tail properties. It demonstrates that the heavy-tail property belongs to deep neural networks rather than SGD itself.

We presented the power-law spectrum of two-layer FCN on Avila Dataset in Figure 13. It shows that the power-law spectrum of neural networks may also generally exist in non-convolution neural networks trained on a non-image dataset. Particularly, we note that the Avila Dataset has only ten attributes, including intercolumnar distance, upper margin, lower margin, exploitation, row number, modular ratio, interlinear spacing, weight, peak number, and modular ratio/ interlinear spacing. These attributes are essentially different from the pixels in image datasets.

We presented the power-law spectra of ResNet18 on CIFAR-10 in Figure 14. It shows that the power-law spectra hold for ResNet, a representative of the modern neural network architectures, as well as simple CNNs/FCNs. Due to the GPU memory limit, we may only display the top 50 eigenvalues for ResNet18. However, the KS test still supports accepting the power-law hypothesis.

The spectra of LeNet on CIFAR-10 trained via various optimizers are showed in Figure 15. Figures 4 and 16 shows that the slope magnitude $\hat{s}$ closely correlates with the largest Hessian eigenvalue and the Hessian trace.

Figure 17 shows that the small width of neural networks may also break the power-law spectrum like small depth. This also supports that overparameterization or large model capacity is necessary for the power-law spectrum.

We report the spectra of large-batch trained ResNet18 on CIFAR-10 in Figure 18. It indicates that the phase transition behaviors of the spectra with respect to batch size generally exist. However, it seems that Phase II and Phase III merge into one phase for ResNet18 on CIFAR-10.

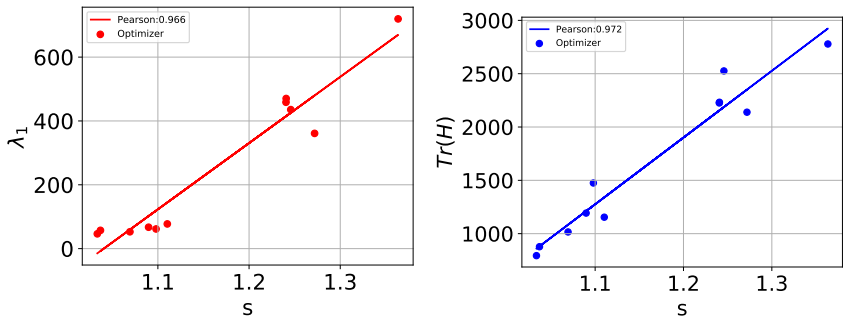


Figure 16: The slope magnitude $\hat{s}$ closely correlates with the largest Hessian eigenvalue and the Hessian trace. Model:LeNet. Dataset: CIFAR-10.

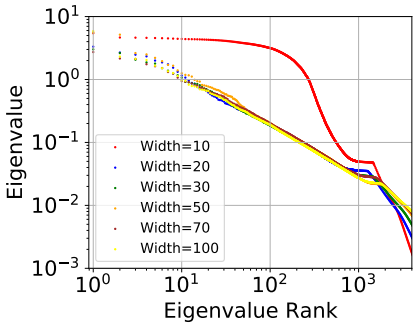


Figure 17: The spectra are not power-law for neural networks with a small width(~ 10), but gradually become more power-law (more straight in the log-log plot) as the width increases. This may also suggest that the power-law spectrum depends on model capacity. Model: Two-layer FCN. Dataset: MNIST.

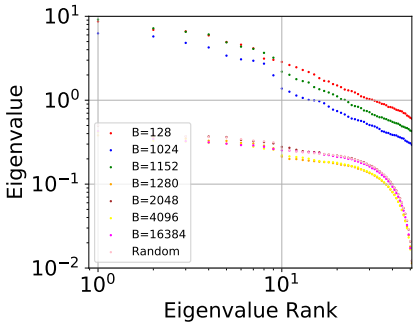


Figure 18: Batch size matters to the spectrum. Model:ResNet-18. Dataset: CIFAR-10. The sharp phase transition occurs in $1152 < B < 1280$.

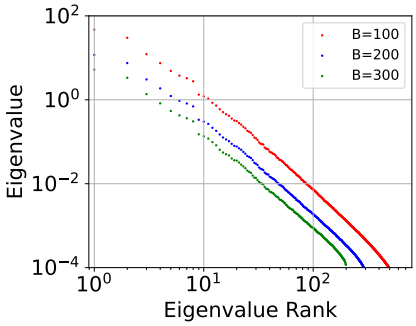


Figure 19: The power-law spectrum of gradient noise covariance exists in deep learning for various batch sizes. Model: Fully Connected Network(FCN). Dataset: MNIST.

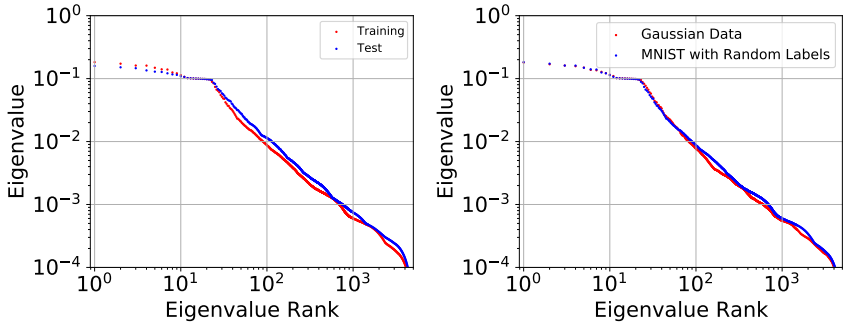


Figure 20: (a) The spectrum of LeNet on (training and test) MNIST with randomly shuffled labels. (b) The spectrum of LeNet on Gaussian data with randomly shuffled labels is highly similar to that MNIST with randomly shuffled labels.

The heavy-tail property of SGD has been a hot and arguable topic recently (Simsekli et al., 2019; Panigrahi et al., 2019; Gurbuzbalaban et al., 2021; Hodgkinson & Mahoney, 2021; Xie et al., 2021a; Li et al., 2021). Note that the power-law distribution is one of the most common heavy-tail distributions in the real world. We argue that the arguable heavy-tail property of SGD may depend on the power-law Hessian spectrum rather than SGD itself, as gradient noise covariance critically depends on the Hessian. We present the power-law spectra of gradient noise covariance in Figure 19.

We presented the spectrum of learning with clean labels and random labels in Figure 20. The number of top outliers obviously increases, because random labels make the dataset more complex. However, even if the pairs of instances and labels have little knowledge, we still observe the power-law spectrum after the dozens of top outliers. This may suggest that, even if the labels are random, neural networks can still learn useful knowledge from the instances only.