Deep Learning, Correlations, and the Statistics of Natural Images

Robert W. Batterman
University of Pittsburgh
Department of Philosophy
November, 22 2024
IPAM
Modeling Multi-Scale Collective Intelligences

Figure: Model Organisms: Quinn and Devi

- Deep Neural Networks (DNNs) are remarkably successful: They generalize well when trained on various datasets, predict 3D protein folding structures, ...
- Nevertheless, there remain significant questions as to how: Issues about overfitting and generalization, among others things.
- A number of authors have suggested or implied that one might explain (some) successes by arguing that DNNs (at least for certain tasks) are implementing Renormalization Group (RG)-like operations.

- Here I want to explore and go beyond some aspects of these suggestions.
- While I believe claims of an exact mapping between DNNs and the RG are likely overblown, certain DNNs can be understood as extracting relevant correlational information that is "hidden" in the input data.

Broad reasons for thinking there is at least some kind of connection between RG and DNNs:

- The RG, e.g., allows one to extract information concerning continuum scale (e.g., critical) behavior from all the (largely) irrelevant gory details about the molecular makeup of different fluids.
- DNNs, e.g., apparently allow one to extract information relevant to classifying an image as that of a dog from all of the (largely) irrelevant gory details present in the pixels making up the image.

- I want to suggest that these procedures are successful, to the extent that they are, because there is genuine correlational information hidden in the gory details.
- I also think we can say something about the nature of the correlational information and where it comes from.
- These correlations reflect the fact that the world exhibits various kinds of scale invariances.
- The correlations inhabit a scale in between the microlevel and the macro-level.
- This may also help to explain the hierarchical nature—the depth—of the DNNs.

- Ruderman and Bialek (1994) took a series of photographs in a state park in New Jersey.
- The photos were primarily of trees, rocks, and a stream. They measured 256 by 256 pixels and corresponded to 15 degrees in visual angle. The data they collected were the logarithm of each pixel's luminance. (Ruderman, 1997, p. 3386).

- The data showed scaling "in the power spectrum of the form:

$$S(k) = \frac{A}{k^{2-\eta}}, \tag{1}$$

  with k being the spatial frequency, A is a constant representing the overall contrast power in the images . . . ."
- For their data the "anomalous" exponent η had a value of 0.19.

How can we understand this power law scaling?

- One could imagine forming blocks of pixels (in analogy with block spins in a real-space renormalization scheme). We would see the same statistical structure in the pixel-blocked images after appropriate renormalization. Kadanoff (2013).

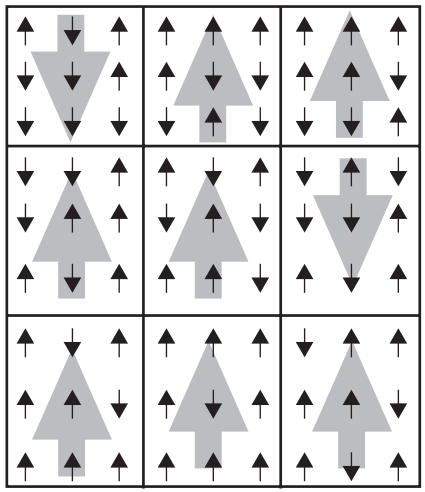


Figure: Blocking and averaging to yield a new (coarse-grained) effective system (Kadanoff, 2013, p. 172)

Ruderman and Bialek actually do this pixel blocking. They plot the contrast, ϕ, of the images (normalized to unit variance) averaged over *NxN* pixel blocks for ($N = 1, 2, 4, \ldots, 32$). Each such plot superposes on the same (non-Gaussian) distribution:

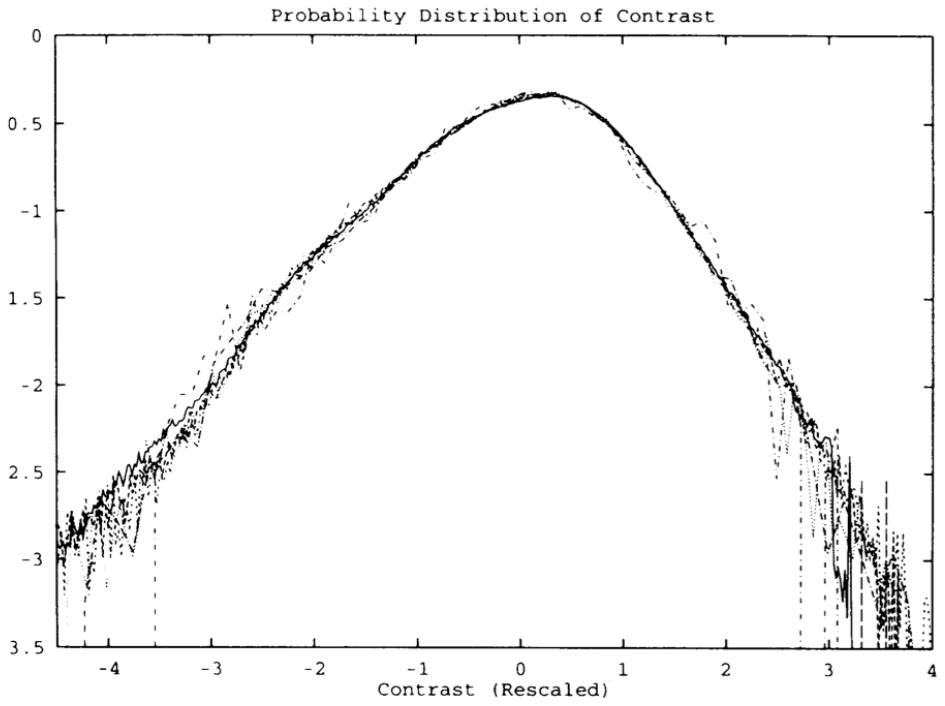


Figure: Distribution of Contrast.

- Another indication of the robustness of the statistical structure in the natural images is shown by implementing certain quite radical database recalibrations of the data.
- Changing the gray scale images in the database to black and white produces a new data set with virtually the same statistics.

That the process of geological formation of hillsides and valleys, or the structure of forests due to the succession of flora, can exhibit scaling through their images is perhaps not altogether surprising. . . . It is striking, however, that the natural image datasets in which scaling was found are all quite different. No two sets of pictures were even from the same environment. (Ruderman, 1997, pp. 3385–3386)

- The scaling result (1) is a function of the spatial **frequency** k.
- Working in the frequency domain is not the best way to understand what properties of natural images are responsible for scaling. In the Fourier/frequency domain, objects are spread out and superpose over many frequency bands.

Ruderman reformulates the results in the spatial domain, introducing a pixel "difference function" that will allow him to "define" objects statistically.

$$D(x) = \langle\, |\phi(0) - \phi(x)|^2 \,\rangle, \tag{2}$$

(x is measured in degrees of visual angle.)

This is a kind of expected variance. Pixels in different objects have a larger mean squared difference in luminance than those that belong to the same object.

Expanding the square in the difference function (2), Ruderman shows it takes the form

$$D(x) = D_1 - D_2 x^{-\eta}.$$

- The values for D_1 and D_2 were determined by randomly selecting 10^5 pixel point pairs from the images and tabulating their joint statistics.
- This yielded a value $\eta = 0.19$ in accordance with the anomalous dimension from the power spectrum.

How can this be explained? That is, what is responsible for the form of the pixel difference function $D(x)$? Specifically, what explains the value of the scaling exponent η?

- A simple but important observation: Each point pair of pixels from an image can either belong to the same object or to two different objects.
- We need to determine the statistics—the probabilities that such point pairs are in the same object or not.
- Another important feature of natural images is that objects in the images can occlude one another.
- Ruderman constructs a simple model of image generation that allows for this occlusion and that lets us identify statistically independent image components as distinct "objects." Thus, objects are defined **statistically and not semantically**.

The model:

> *Imagine walking on an infinite image plane. At a random location you blindly select from a number of choices an infinitesimally thin cardboard "cut-out" of some shape. You paint it a gray tone chosen from a distribution, and then drop it on the ground. This done, you continue to another random location and repeat the process. (Ruderman, 1997, p. 3392)*

- For this model the correlation function is:

$$C(x) = C_0 P_{\text{SAME}}, \tag{3}$$

where P_{SAME} is the probability that a given point pair separated by a distance x belong to the same object and C_0 is the constant correlation within objects. (Ruderman, 1997, p. 3392)

- A power law in the correlation function means a power law in the spectrum. To show this we need to show that P_{SAME} is itself a power law. (Ruderman, 1997, p. 3393)

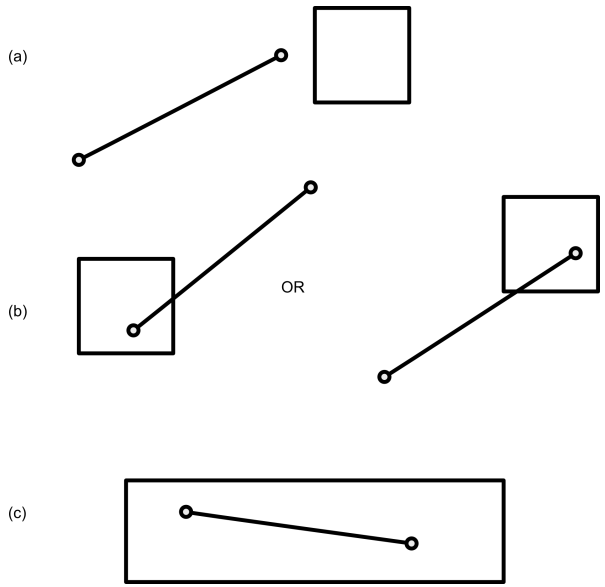


Figure: Randomly throw Line Segments of length x on a plane with rectangular objects. (Ruderman, 1997, p. 3393)
(a) yields $p_0(x)$; (b) yields $p_1(x)$; (c) yields $p_2(x)$.

We get

$$P_{\text{SAME}}(x) = \frac{p_2(x)}{p_1(x) + p_2(x)}, \tag{4}$$

where $p_1(x)$ and $p_2(x)$ are determined by examining figure 4.

Ruderman concludes that

> *. . . the scaling of inter-object probability follows directly from the scaling of apparent object sizes. In images of the real world this apparent size (in degrees) depends on an object's actual size as well as its distance from the observer. The overall distribution of apparent object size is thus a function of the distributions of object sizes and that of their distances. (Ruderman, 1997, p. 3393)*

Note that higher order (N-point) correlation functions can be easily discovered as well. What is the probability that three pixels in an image lie in the same object? Four pixels? . . .

Figure: Throwing Lines, Triangles and . . . . Batterman (2021)

Ruderman's scheme for determining two-point correlation functions between pixels in the images is an instance of a widely applicable multi-scale methodology for understanding the behavior of many-body systems in condensed matter physics and materials science.

- This methodology was promoted by Leo Kadanoff and Paul Martin, and is sometimes referred to as a set of hydrodynamic or correlation functions methods.
- In these many-body systems, some of the most important quantities that appear at continuum scales are so-called "order parameters" and "material parameters."
- Examples, respectively, include the net magnetization of a ferromagnet and the viscosity of a fluid.

I've argued that values for these parameters are actually coding for **correlational structures** at mesoscales in between the microscopic (atomic/molecular) and the continuum. Batterman (2021)

- One can see this by noting, e.g., that the net magnetization M is defined as the difference in the densities of up-spins ($\rho_\uparrow$) and downspins ($\rho_\downarrow$).

$$M(\mathbf{r}, t) = |\rho_\uparrow(\mathbf{r}, t) - \rho_\downarrow(\mathbf{r}, t)|,$$

  at spatiotemporal points (**r**,t).
- Thus, material and order parameters are **defined** at mesoscales—one cannot "see" densities at atomic or lattice scales.

- The key take-away is that the scaling structure in natural images is a function of certain correlations between pixel values.
- The power law scale invariance in natural images is a function of that correlational information.
- As with many-body systems, this information is discoverable at scales between that of pixels and the entire image.
- It is also important that the scaling of inter-object probability is a function of the **actual** distribution of object sizes and their distances.
- I am going to suggest that the effectiveness of DNNs in image recognition (among other tasks) depends upon the existence of that correlational information—correlational information that reflects features of the **real world**.

An influential paper by H. Lin, M. Tegmark, and D. Rolnick entitled "Why Does Deep and Cheap Learning Work So Well?," aims to show how the success of deep learning depends not only on the mathematics of neural networks but also on certain facts about the world.

- They raise a puzzle: "*How can neural networks approximate functions well in practice, when the set of possible functions is exponentially larger than the set of practically possible networks?*". (Lin et al., 2017, p. 1225)

- This question arises because even networks with only one hidden layer are known to be universal function approximators.
- Given a sufficient number of hidden units any smooth function can be approximated to any accuracy with just a single hidden layer.
- Lin et al., give a quick estimate that demonstrates that networks of "feasible size" however cannot do this. "There are 2^{2^n} different Boolean functions of n variables, so a network implementing a generic function in this class requires at least 2^n bits to describe, *i.e.*, more bits that there are atoms in our universe if $n > 260$. (Lin et al., 2017, p. 1228)

- The space of all functions is enormous.
- Despite this, neural networks of "feasible size" (read "actually implementable) have been extremely successful.
- Lin et al. argue that for "physics reasons" scientists/physicists typically only care about a very small subset of the space of functions.
- These, they say, are functions (Hamiltonians) that have low polynomial order, that exhibit certain symmetries, and that involve local interactions.
- The kind of functions we want to approximate are extremely far from being random. In effect, they argue that one reason DNNs work well is because the space of functions we actually care about is extremely small in the space of all functions.

- There is something right about this. But, it cannot be the whole story: It isn't much of an explanation.
- Our interests can (should) only be part of the reason DNNs work.
- Not all Deep Learning is aimed at problems in physics. What is the Hamiltonian for an image of a dog?
- Of real interest is how DNNs actually find the functions that work—the functions that correctly recognize objects at the scale of dogs—given input at the scale of pixels.
- The explanation for this must appeal to actual facts about the world—the scale invariances—and to the means by which the DNNs find functions that detect the correlations that yield those invariances.

- It is well-known that the training of DNNs takes place using a variety of datasets. In the context of image recognition these include:
  1. MNIST—A large database of handwritten numbers
  2. FMNIST—An MNIST-like database of labeled fashion images.
  3. CIFAR10—A very large database of labeled images from 10 classes representing airplanes, birds, cars, cats, deer, dogs, frogs, horses, ships, and trucks.
  4. IMAGENET—A huge database containing more than 14 million labeled images from more than 20,000 classes or categories.
  5. And many others.

There have been empirical studies of the statistics of these datasets using the theory of Random Matrices (RMT).

- These involve the study of eigenvalue spectra of matrices representing samples from the datasets

$$\Sigma_M = \frac{1}{M} X^T X,$$

  where $X \in \mathbb{R}^{d \times M}$ and d is the dimension of the image vectors and M is the number of samples. Levi and Oz (2024)
- Σ_M is an empirical covariance (Gram) matrix.

- Empirical investigation shows that the bulk of the eigenvalues for various datasets represent "the correlational structure of different features amongst themselves, . . . " and that these "decay as a power law $\lambda_i \propto i^{-1-\alpha}$." (Levi and Oz, 2024, p. 2)
- Levi and Oz explicitly reference Ruderman's scaling law in this context.
- Remarkably, the various datasets mentioned earlier all exhibit the **same** power law behavior.

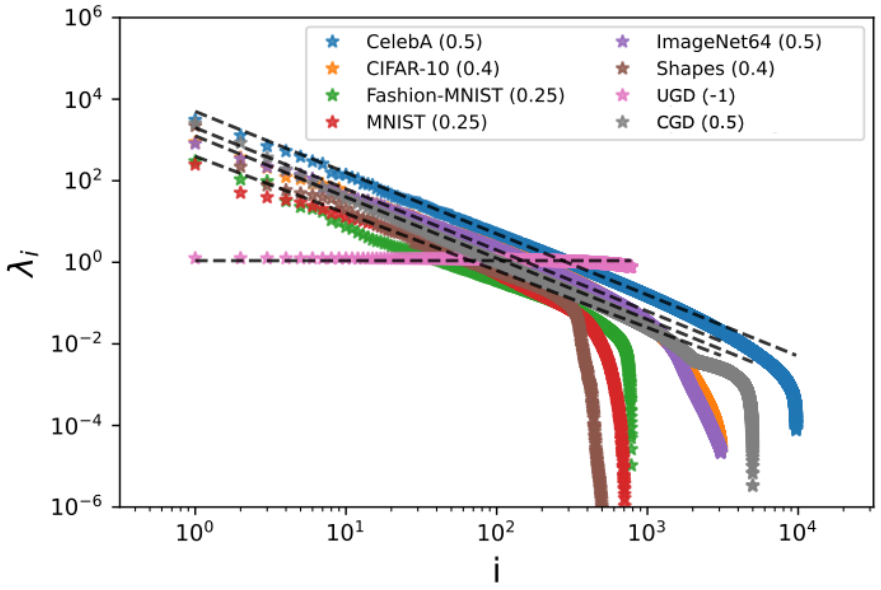


Figure: Scree Plots: Scaling Behavior of Σ_M for Various Datasets (Levi and Oz, 2024, p. 4)

- Eigenvalue bulk $\lambda_i \propto i^{-1-\alpha}$.
- All real-world datasets have $\alpha \leq 1/2$.
- The value of α reflects the strength of the correlations in the covariance matrices.

The fact that images of the world present with certain statistical structure is critical to understanding how DNNs are actually able to generalize.

- Here I want to present results of empirical studies on actual DNNs that have been trained on various datasets:
- "Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning" Martin and Mahoney (2018)

- Martin and Mahoney give evidence that the Empirical Spectral Density (ESD), $\rho_N(\lambda)$ of the correlation matrices exhibits statistics from an RMT universality class of **Heavy-Tailed distributions**.
- These are power-law distributions and they often are the mark of complex systems.

Martin and Mahoney represent the energy landscape (or optimization function) of a "typical" DNN having L layers with activation functions $h_l(\cdot)$, weight matrices per layer $\boldsymbol{W}_l$, and biases $\boldsymbol{b}_l$ as:

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

- They study the weight matrices $\boldsymbol{W}_l$ **before**, **during**, and **after** training on various databases and for a wide range of actual DNN models.
- Specifically, they analyze the Empirical Spectral Density, $\rho_N(\lambda)$, of the correlation matrix $\boldsymbol{X} = \boldsymbol{W}^T\boldsymbol{W}$ associated with the layer weight matrix $\boldsymbol{W}$.

- Random Matrix theory provides Law of Large Numbers-like and Central Limit Theorem-like results for matrices.
- RMT yields unique results for both square and rectangular matrices.
- However, in DNNs square weight matrices are rare. Typically the number of parameters (N) is greater than the number of examples (M).
- Much work in RMT has focused on a class of matrices that are members of the **universality class of Gaussian distributions**.

- Correlation matrices in this class have Spectral Density functions

$$\rho_N(\lambda) := \frac{1}{N}\sum_{i=1}^{M}\delta(\lambda-\lambda_i)$$

that in the limit $N \to \infty$ (with aspect ratio $Q = N/M \geq 1$ fixed), takes the form of the Marčhenko-Pastur (**MP**) distribution:

$$\lim_{N\to\infty}\rho_N(\lambda) = \begin{cases} \frac{Q}{2\pi\sigma^2_{mp}}\frac{\sqrt{(\lambda^+-\lambda)(\lambda-\lambda^-)}}{\lambda}, & \text{if } \lambda\in[\lambda^-,\lambda^+] \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

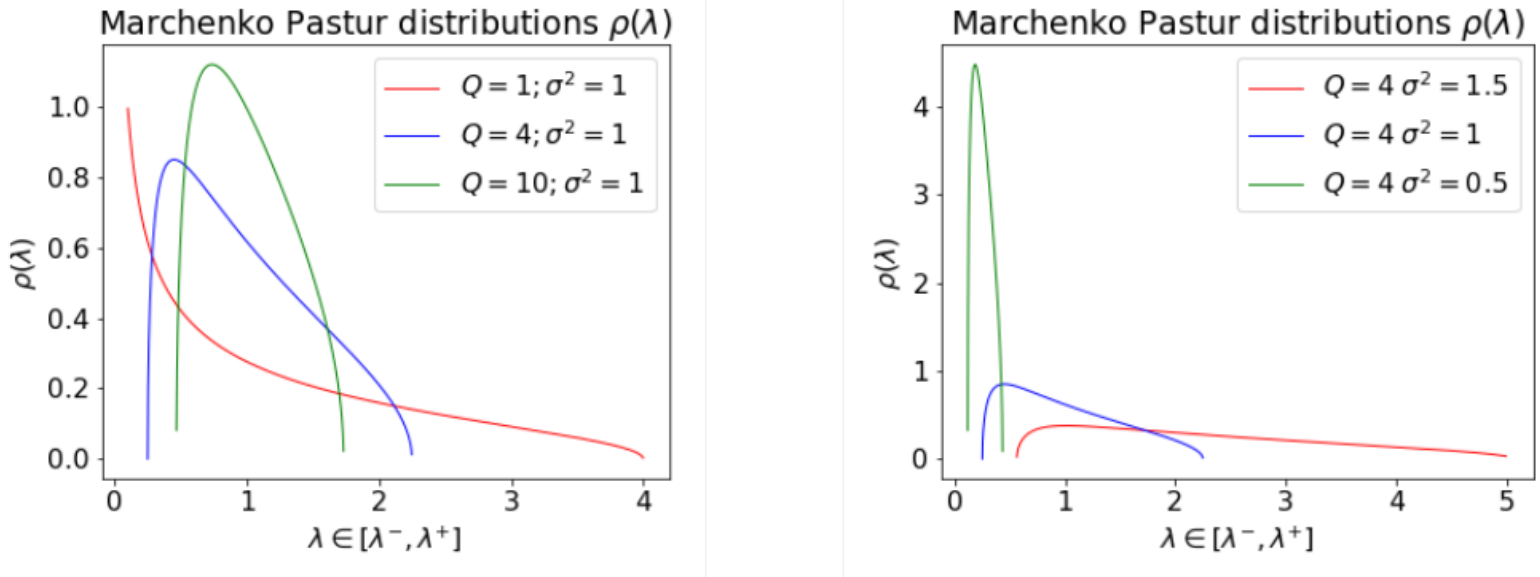
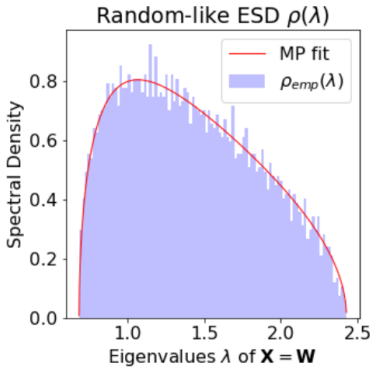


Figure: Left: Different Aspect Ratios. Right: Different Variance Parameters.(Martin and Mahoney, 2018, p. 14)

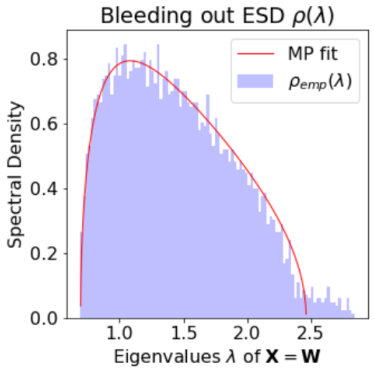
- These, in effect, are the RMT analogs of various Gaussian/normal distributions in ordinary probability theory.

Martin and Mahoney **empirically** investigate the statistics of weight matrices for fully connected layers in a number of **trained** of state-of-the-art DNNs.

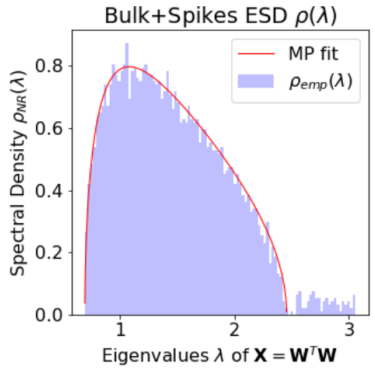
- They observe "profound deviations from traditional [MP-based] RMT." And they find that these DNNs "are reminiscent of strongly-correlated disordered systems that exhibit Heavy-Tailed behavior."
- Most importantly, they argue that the training process for these DNNs "itself engineers a form of implicit *Self Regularization* into the trained model." (Martin and Mahoney, 2018, p. 29)



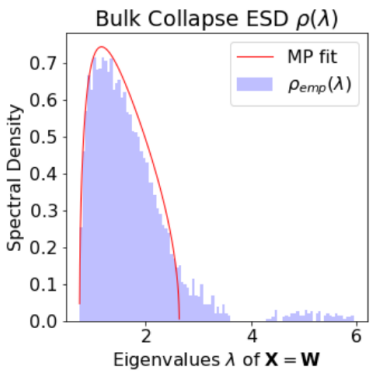
(a) Random–like.



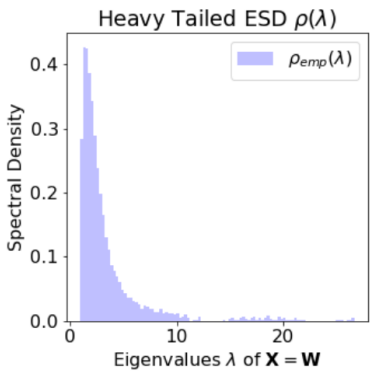
(b) Bleeding–out .



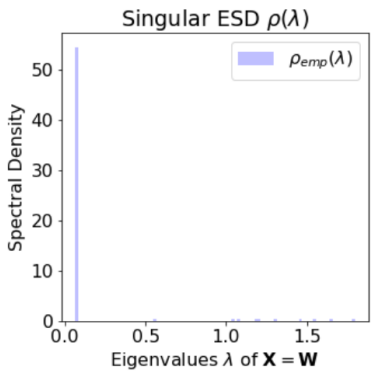
(c) Bulk+Spikes.



(d) Bulk–decay.



(e) Heavy–Tailed.



(f) Rank–collapse.

Figure: Taxonomy of Trained Models. Changing RMT statistics for Weight Matrix Spectral Densities (Martin and Mahoney, 2018, p.32)

- The figure exhibits the evolution of the statistics of the Empirical Spectral Density ($\rho_N(\lambda)$) of the correlation matrices $\boldsymbol{X} = \boldsymbol{W}^T \boldsymbol{W}$ associated with the layer weight matrix $\boldsymbol{W}$.
- These ESDs evolve under training using Stochastic Gradient Descent (SGD) from **random-like distributions** (with good Marčhenko-Pastur fit) **associated with random initialization at the start of training**, to Heavy-Tailed distributions that correspond to strong correlations in $\boldsymbol{W_l}$ for layers $\boldsymbol{l}$ **at the end of training**.
- The idea is that one can model the ESDs of trained DNNs with Heavy-Tailed distributions using RMT.

Martin and Mahoney note that "[f]or DNNs, these correlations arise in the weight matrices during Backprop training [SGD] (at least when training on data of reasonable quality). That is, the weight matrices "learn" the correlations in the data." (Martin and Mahoney, 2018, p.29)

- This, if true, and given the robust statistical power law behavior of various datasets, may help to explain (some of) the remarkable successes of DNNs.

- I hope to have motivated the idea that the success of DNNs at certain tasks can be understood in terms of their ability to discover correlational structures that are present in real world data.
- It is remarkable (at least to me) that the various datasets used to train DNNs all exhibit scaling behavior similar to one another, and similar to the behavior Ruderman's investigations uncovered.
- One way to think about this is in analogy with the universal behavior of critical phenomena.

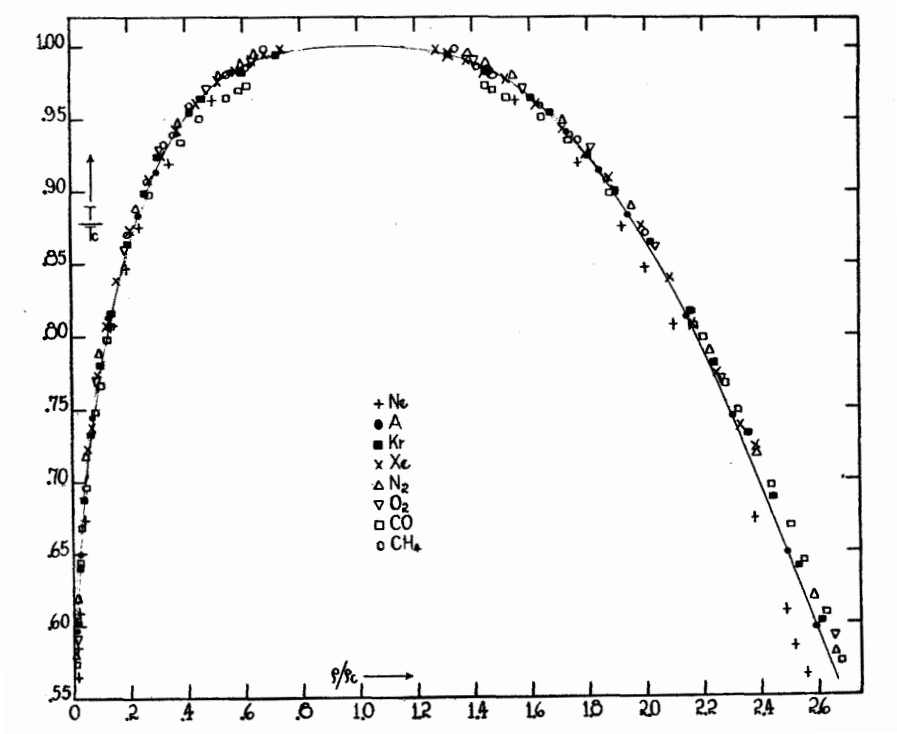


Figure: Universality of Critical Phenomena (Guggenheim, 1945)

- Just as in the figure from Guggenheim, the different datasets upon which DNNs get trained all exhibit the same scaling law behavior.
- This is true despite having widely differing details at "fundamental" pixel scales. That is, despite being datasets of very different images.

- Much of the literature on DNNS (specifically DCNNs) focus on their ability to serve as **feature** detectors—features present in the images they receive as input.
- They are said to detect low level features like edges at early levels in their architectures.
- Then in deeper layers they supposedly manage to combine the lower level features into more "abstract" features that enable them to "recognize/learn" that certain images are images of dogs.
- I think that this talk of "feature detection" is somewhat misleading in that it seems to imply that the DNNs are recognizing certain aspects of **semantic** categories.

- Recall that Ruderman's analysis explicitly avoided a semantic understanding of objects appearing in images.
- Rather than identifying objects in images semantically, as the term "feature detection" suggests, Ruderman argued that we should think of objects as statistically defined.
- In doing this, he was led to determine the probabilities that two pixels separated by some angular distance x belong to the same "object."
- As he says, "[t]he notion of statistically correlated and uncorrelated regions within images corresponding to objects provides a simple, robust path to scale invariance, as long as those objects appear in all sizes according to a power-law distribution." (Ruderman, 1997, p. 3386)

Ruderman's conclusion was that

> *the scaling of inter-object probability follows directly from the scaling of apparent object sizes. In images of the real world this apparent size (in degrees) depends on an object's actual size as well as its distance from the observer.* **The overall distribution of apparent object size is thus a function of the distributions of object sizes and that of their distances.**

Thus, real-world mesoscale structure determines the correlational structure among pixels in images. And this is what ultimately allows the DNNs to correctly characterize the images.

- As Ruderman's argument demonstrates, one can determine the image scaling laws by constructing/finding 2-point correlation functions.
- But this is clearly not sufficient for "deciding" whether or not the image is that of a dog.
- I conjecture that higher order (N-point) correlation functions are required for image recognition.
- It is here that invoking a strong analogy between the RG and DNNs may be fruitful.

A paper by Mehta and Schwab (2014) suggests that there is an exact mapping between RG and deep learning based on stacked restricted Boltzmann machines.

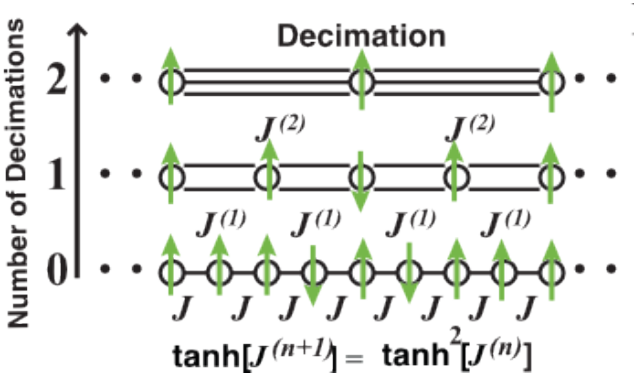


Figure: RG Decimation (Mehta and Schwab, 2014, p. 5)

- Consider a one-dimensional Ising chain—spins on in a line with a fixed spacing between them. There is a coupling between neighboring spins denoted by J.
- Decimating (killing off) every other spin yields a crude coarse-graining analogous to the **block spin** formation considered earlier.
- As before, there are new couplings, $J^{(i)}$s, between the remaining spins.

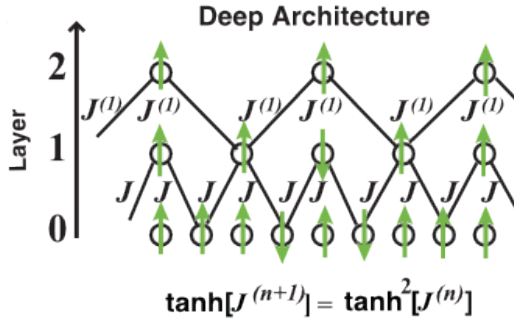


Figure: DNN Analog and "Hidden" Spins (Mehta and Schwab, 2014, p. 5)

- In this figure the $J^{(i)}$s are so-called "**hidden**" spins. (Mehta and Schwab, 2014, p. 5)
- They code for correlations between the spins (or neuron weights) in the previous layer of the neural net.

- I suggest that we think of the blocking/decimation RG scheme as **actively** creating block spins, thereby eliminating microscopic degrees of freedom that are irrelevant for upper, continuum scale physics.
- The DNN, on the other hand, **passively** reveals correlations that are hidden in the microscopic details but that only emerge or become visible at higher scales.
- My suggestion is that the higher layers in the DNNs can be seen as revealing or finding 3-point, 4-point . . . correlation functions present in the pixelated input.
- If this is correct, then it does indeed seem that DNNs are, in effect, reconstructing a "continuum level field" that is the image of a dog.

Finally, there is evidence that DNNs that have been trained on one dataset require considerably less training to become successful image classifiers on other datasets that they have not seen. Why? After all, the datasets are of different images.

- Possibly because they have already achieved a kind of **alignment** with the ground truth function (read robust power law scaling) that is present in all of the datasets. Wei et al. (2022)
- So it shouldn't be surprising that the DNNs port well to other datasets.
- They already have the weight parameter structure corresponding to the (same) statistics in the new datasets.

Thank You!

And, thanks to Katie Creel, Conny Knieling, Sameera Singh, Porter Williams, and Jim Woodward.

References

Robert W. Batterman. *A Middle Way: A Non-Fundamental Approach to Many-Body Physics*. Oxford University Press, 2021.

E. A. Guggenheim. The principle of corresponding states. *The Journal of Chemical Physics*, 13(7):253–261, 1945.

Leo P. Kadanoff. Theories of matter: Infinities and renormalization. In Robert W. Batterman, editor, *The Oxford Handbook of Philosophy of Physics*, chapter Four, pages 141–188. Oxford University Press, 2013.

Noam Levi and Yaron Oz. The underlying scaling laws and universal structure of complex datasets. *arXiv:2306.14975v3*, 2024.

Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap lerning work so well? *Journal of Statistical Physics*, 168: 1223–1247, 2017.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *CoRR*, abs/1810.01075, 2018.

Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning, 2014.

Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997.

Daniel L. Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, pages 23549–23588. PMLR, 2022.