A Data Scaling Law of a Manifold's Resolution

Abiel J. Kim

April 2025

Abstract

It has been empirically observed that neural network performance generally assumes the power law formulation for the scaling of its training dataset. The experimental evidence is compelling, but the theoretical frontier remains exploratory with respect to a mathematical origin of the observed power law. This paper introduces a mathematical framework of the geometric kind that enables the emergence of a bounding of the data scaling law. The mathematical framework is predicated on the manifold conjecture and interprets the scaling of a dataset as a finer approximation to the true data manifold space. The equations indicate that model loss, L, indeed scales as a power law with $L \propto D^{-1/d}$ for the data manifold's intrinsic dimensionality, d.

1 Introduction

As per the question of the neural scaling laws posed by researchers at OpenAI, Kaplan et al. (2020) proposed the empirically observed power law formulations for model capacity (N), dataset size (D), and compute (C).

And as discussed in lecture, we saw the empirical literature surrounding the proposed power law formulation for the scaling of D. We also discussed the idea that there does not yet exist a universally accepted theoretical solution regarding the mathematical origin of such observed power laws.

The objective of this paper is to discover a theoretical upper bound for the data scaling law from first principles which has been empirically fitted as a power law formulation. The implication of such a discovery would provide either confirming or disconfirming evidence for the fitting of the power law with respect to data scaling.

In order to achieve this, I will provide firstly a geometric intuition of my framework. Then, I will provide a more detailed mathematical derivation that leverages some of the mathematical properties of the geometric framework which leverages the manifold conjecture. Note that the manifold conjecture is a mainstream approach in terms of theoretical analysis in the machine learning field.

2 Related Works

Although I did not find any paper that matches my work exactly, the closest work that I could find (with the help of Professor Vincent) is a paper by Bahri et al. (2024) which appears to take a similar conceptual approach that I did with respect to the notion of a data manifold's resolution.

However, the key difference my work and that of Bahri et al. (2024) is the degree of theory in the derivation of the data scaling law. While my work here is purely theoretical i.e. model agnostic, a standalone purely mathematical proof- the work done by Bahri et al. (2024) appears to mix both empirical validation and theoretical work together to derive a data scaling law by leveraging prior work and empirical argumentation.

Ultimately, I did not find any prior works that matches the work that I've done. Though, there is plenty of overlap as my work ultimately leverages the manifold conjecture which is a widely accepted approach in the theoretical analysis of deep learning.

My work was also inspired by that of Sharma et al. (2020) whom took a similar approach to the derivation of the model scaling law for N. Their manifold based approach resonated deeply with me as it matched my original intuition for how deep learning may have deep mathematical ties in the loose context of manifold learning.

3 Manifold Resolution Theorem (Methods and Results)

3.1 The Geometric Intuition

This geometric framework assumes the manifold conjecture, in which a given data set corresponds to a sampling of an underlying d-manifold in multidimensional space. i.e. the linear regression architecture assumes that the dataset is a sampling of a linear n-manifold in $n+1$ space with the trivial case of a line manifold embedded in 2 dimensional space. The decision boundaries correspond to hyperplanes that subdivide the n-manifold, and it is the model's objective to distinguish between these hypersubspaces by minimizing the loss between prediction and truth.

The natural extension for a more complex dataset is the definition of a more complicated d-manifold geometry. The feature space may no longer be linearly correlated, and hence the underlying manifold structure may assume a highly irregular, nonlinear structure. *We are permitted to interpret the dataset as a discrete sampling of the underlying lower-dimensional d-manifold* with the underlying manifold being a continuously defined structure. The objective of the deep neural network should then be the subdivision of the d-manifold into hypercubic regions that correspond to class membership mappings.

In the limit, as the size of the dataset, D, scales, observe that a better approximation to the underlying d-manifold structure is attained. In other

words, as D scales, the resolution of the d-manifold structure increases and the structure clarifies. Therefore, realize that if D approaches infinity, then a perfect representation of the d-manifold is achieved.

3.2 Definitions and Notation

Let us define the input space, $\mathcal{X} \subseteq \mathbb{R}^N$, of dimension $dim(\mathcal{X}) = N$. Embedded within $\mathcal{X}$ there exists the d-manifold structure, $\mathcal{M} \subseteq \mathbb{R}^N$, of intrinsic dimensionality $dim(\mathcal{M}) = d$ that is smooth and compact with $d << N$. The consequence of compactness is the assertion of a finite volume $V_{\mathcal{M}} < \infty$ that the manifold inhabits. Further, we assume that the dataset $\{x_1, x_2, \ldots, x_D\} \in \mathcal{M}$ gets sampled from the surface of the data manifold at i.i.d. with uniform probability $p(x)$.

Next, we shall define the hypothesis class, H, of Lipschitz functions that maps a sample on the surface of $\mathcal{M}$ to a real number expressed as $f : \mathcal{M} \to \mathbb{R}$ and $|f(x)-f(z)| \leq L|x-z|, \forall f \in H$ for a real positive constant L. The predictive function $\hat{f} \in H : \mathcal{M} \to \mathbb{R}$ corresponds to our learned model mapping. The true function $f^* : \mathcal{M} \to \mathbb{R}$ represents the theoretically perfect mapping which may exist outside of the hypothesis class such that $f^* \in H$ or $f^* \notin H$. However, we also assume that the true function is Lipschitz, thus $f^* : |f^*(x) - f^*(z)| \leq K|x - z|$ for a real positive constant K. The Lipschitz constraints imposed upon H and f^* reduce the set of all possible functions to those that do not oscillate rapidly between arbitrary pairs of neighboring data points upon the surface of the data manifold.

The true risk $R(f)$ for some arbitrary $f \in H$ is the expected MSE between $f \in H$ and the true function f^*. If we assume that data is sampled i.i.d. from the d-manifold surface at uniform probability $p(x)$ then we formulate true risk as the integral $R(f) = E[(f(x) - f^*(x))^2] = \int_{\mathcal{M}} (f(x) - f^*(x))^2 p(x) dV_{\mathcal{M}}$ for some $f \in H$ where x lies on the surface of $\mathcal{M}$. The empirical risk, $\hat{R}_D(f)$ for some $f \in H$, is the average MSE between f and the true function f^* over D data points. This is equivalent to the training error and can be simply expressed as $\hat{R}_D(f) = 1/D \sum_{i \leq D} (f(x) - f^*(x))^2$.

Correspondingly, the true minimizer $f_F^* \in H$ is the optimal function with minimum true risk such that $f_H^* = argmin_{\forall f \in H} R(f)$. Then, we shall define the empirical minimizer $\hat{f}_D \in H$ that corresponds to the optimal function with minimum empirical risk over D discrete points $\{x_1, x_2, ..., x_D\} \in \mathcal{M}$ such that $\hat{f}_D = argmin_{\forall f \in H} \hat{R}_D(f)$.

If f_H^* is the best approximation from H to f^* over the population dataset and $\hat{f}_D$ is the best approximation from H to f^* over D sampled data points, then we must discover and bound the excess risk $R(\hat{f}_D) - R(f_H^*)$ as $D \to \infty$ from first principles.

3.3 Reiteration of Key Assumptions

Assumption 1: $\mathcal{M}$ is smooth and compact i.e. $\mathcal{M}$ is differentiable and bounding of a finite volume. Assumption 2: The dataset $\{x_1, x_2, ..., x_D\} \in \mathcal{M}$ is distributed uniformly across the data manifold. When sampling, we assume points are taken with a uniform probability distribution at i.i.d. Assumption 3: The Lipschitz hypothesis class H comprises smooth, non-jagged function surfaces. The true function f^* is also Lipschitz.

3.4 Modeling the Data Manifold Resolution

As D increases, data points inhabit the data manifold's volume at increasing resolution. If $V_{\mathcal{M}}$ is finite and we have D sample points that are uniformly distributed across $V_{\mathcal{M}}$ then each data point inhabits a region of $V_{\mathcal{M}}/D$ space on average.

The volume of a data point in $\mathcal{M}$ can be modeled as the volume of a d-ball in d space, $R^d V_d$, where V_d is the volume of the unit d-ball and R^d is its radius. For instance: If $d = 2$ then $V_d = \pi$, if $d = 3$ then $V_d = 4\pi/3$, and so forth.

It therefore follows that we can express the average volume that each data point occupies as $R^d V_d = V_{\mathcal{M}}/D$. We are permitted to express it in this way since we assume the uniform distribution of data points across the manifold. Solve for radius to find $R = (V_{\mathcal{M}}/V_d)^{1/d} D^{-1/d}$.

Furthermore, assumption 2 permits us to interpret R as an approximation to the typical radius. Therefore, we may use R to approximate the average distance between neighboring data points upon the surface of the manifold. Finally, express typical radius as $R = \phi D^{-1/d}$ which gives us a measure of the manifold's resolution.

Note that we have modeled the volume of the inhabited region by each data point as a hypersphere and not some other geometry such as a hypercube. This is justified because whether the region of each data point is modeled as a hypercube or any other arbitrary shape in d-space, the formulation for R typically yields the same structure: $D^{-1/d}$ multiplied by some constant. It therefore follows that as $D \to \infty$, the average-volume approximations modeled with different hyper geometries converge to the same typical radius. The decision of a hypersphere is chosen more for its convenience.

3.5 Upper Bound of the Excess Risk

Rewrite excess risk $R(\hat{f}_D) - R(f^*_H) = (R(\hat{f}_D) - \hat{R}_D(\hat{f}_D)) + (\hat{R}_D(\hat{f}_D) - R(f^*_H))$. And since $\hat{R}_D(f^*) - \hat{R}_D(f^*) = 0$, mutate the second difference to rewrite excess risk again as $R(\hat{f}_D) - R(f^*_H) = (R(\hat{f}_D) - \hat{R}_D(\hat{f}_D)) + (\hat{R}_D(\hat{f}_D) - \hat{R}_D(f^*_H)) + (\hat{R}_D(f^*_H) - R(f^*_H))$.

Observe the middle difference and realize that $(\hat{R}_D(\hat{f}_D) - \hat{R}_D(f^*_H)) \leq 0$ since by definition, $\hat{f}_D$ minimizes $\hat{R}_D(f)$, $\forall f \in H$ so either the empirical risk of f^*_H matches it or yields a higher risk output. It therefore follows that $R(\hat{f}_D) -$

$R(f_H^*) = (R(\hat{f}_D) - \hat{R}_D(\hat{f}_D)) + (\hat{R}_D(\hat{f}_D) - \hat{R_D}(f_H^*)) + (\hat{R}_D(f_H^*) - R(f_H^*)) \leq (R(\hat{f}_D) - \hat{R}_D(\hat{f}_D)) + (\hat{R}_D(f_H^*) - R(f_H^*))$.

Now we add two differences of empirical and true risk in the final inequality. Further bound this expression by selecting the minimum difference between true and empirical risk that satisfies a further upper bounding. By transitivity, we find the valid upper bound for empirical risk: $R(\hat{f}_D) - R(f_H^*) \leq (R(\hat{f}_D) - \hat{R}_D(\hat{f}_D)) + (\hat{R}_D(f_H^*) - R(f_H^*)) \leq 2sup_{f \in H}|R(f) - \hat{R}_D(f)|$.

3.6 Connecting the Manifold Resolution and Excess Risk Upper Bound

By MSE and transitivity, we eventually find that $|R(f) - \hat{R}_D(f)| \leq E[|h(x) - h(x_i)|]$ where $h(x) - h(x_i) = (\hat{f}_D(x) - f^*(x))^2 - (\hat{f}_D(x_i) - f^*(x_i))^2$ and $|h(x) - h(x_i)| \leq 2\gamma\xi r$ for constants γ and ξ.

Realize that $\forall x \in \mathcal{M}, \exists x_i : |x - x_i| \leq R$ since D points inhabit the surface of $\mathcal{M}$ at scale R. Thus $|h(x) - h(x_i)| \leq 2\gamma\xi R$. It therefore follows that since $|R(f) - \hat{R}_D(f)| \leq E[|h(x) - h(x_i)|]$ and $E[|h(x) - h(x_i)|] \leq 2\gamma\xi R$ then by transitivity excess risks is bounded by $2\gamma\xi\phi D^{-1/d}$.

Ultimately, the excess risk is bounded by the supremum from section 2.5. By transitivity, we can finally formulate our excess risk bound as $|R(\hat{f}_D) - R(f_H^*)| \leq 2\gamma\xi\phi D^{-1/d}$ and subsume all constants for a simplified expression: $|R(\hat{f}_D) - R_D(f_H^*)| \leq \kappa D^{-1/d}$.

4 Discussions

The proof demonstrates that the excess risk has an upper bound of the power law formulation. More concretely, the equations suggest that excess risk shrinks proportionally to the typical radius, R, as D scales to infinity.

From a geometric perspective, you can imagine individual data points uniformly distributed along the manifold surface. For large D, picture a fine grid-like mesh whose vertices correspond to data points. It is intuitive to imagine that the typical radius between points on the grid shrinks as $D \to \infty$ since $V_{\mathcal{M}}$ is finite from assumption 1. More precisely, this is because the typical radius can be expressed as a power law function of the dataset size and intrinsic dimensionality of the data manifold: $R \propto D^{-1/d}$.

However, we were able to connect excess risk to the typical radius, R, and hence the power law of D, showing that the excess risk which shrinks proportionally to R is also bounded by the power law if at least constrained by our 3 assumptions from section 2.3.

Interestingly, our theorem suggests that the excess risk shrinks slower as the manifold's intrinsic dimensionality, d, increases. This indicates that as the underlying data increases in complexity, excess risk is tougher to reduce which is an intuitive notion. Geometrically, one may interpret it as requiring exponentially more data points to cover higher dimensional space.

5 Connection to Class Themes

This paper connects primarily to the topics discussed in week 7 of Professor Vincent's CMPT 419. In week 7, we discussed the notion of the neural scaling laws, and in particular, the data scaling laws and how we have strong empirical evidence to suggest a power law formulation but that the theoretical frontier remained exploratory.

It was during the discussions in lecture where I was inspired to tackle this problem. Frankly, I didn't expect to find a proof in the manner I did, but I did feel an obligation to explore the topic more deeply as the mathematical and theoretical nature of the topic resonated deeply with me. The philosophical implications of the findings could also host significant discussion as to notions beyond machine learning. For instance, if the nature of "machine intelligence" is constructed in the image of man and that there exists an upper theoretical bound for such computational intelligence, what does that suggest of a theoretical upper bound of the intelligence of man? Does such a bound exist? What about a theoretical upper bound of universal intelligence? Such ideas resonated deeply with me and I felt a predisposition to explore a mathematical origin that may satisfy or answer my questions.

6 Acknowledgments

I would like to thank Professor Nicholas Vincent of Simon Fraser University (SFU) for introducing me to the *neural scaling laws* topic and problem during his teachings of the course *Special Topics in Artificial Intelligence, CMPT 419* where I presented my findings. I also thank the colleagues of Professor Vincent at the University of Illinois Urbana-Champaign (UIUC) from whom I received technical feedback and advice regarding a future course of action.

References

[1] Yasaman Bahri et al. "Explaining neural scaling laws". In: *Proceedings of the National Academy of Sciences* 121.27 (2024), e2311878121. DOI: 10.1073/pnas.2311878121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2311878121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2311878121.

[2] Pawel Gajer and Jacques Ravel. *The Geometry of Machine Learning Models*. 2025. arXiv: 2508.02080 [cs.LG]. URL: https://arxiv.org/abs/2508.02080.

[3] Cédric Gerbelot et al. *Applying statistical learning theory to deep learning*. 2024. arXiv: 2311.15404 [cs.LG]. URL: https://arxiv.org/abs/2311.15404.

[4] Alex Havrilla and Wenjing Liao. *Understanding Scaling Laws with Statistical and Approximation Theory for Transformer Neural Networks on Intrinsically Low-dimensional Data*. 2024. arXiv: 2411.06646 [cs.LG]. URL: https://arxiv.org/abs/2411.06646.

[5] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG]. URL: https://arxiv.org/abs/2001.08361.

[6] Utkarsh Sharma and Jared Kaplan. *A Neural Scaling Law from the Dimension of the Data Manifold*. 2020. arXiv: 2004.10802 [cs.LG]. URL: https://arxiv.org/abs/2004.10802.

[7] Namjoon Suh and Guang Cheng. *A Survey on Statistical Theory of Deep Learning: Approximation, Training Dynamics, and Generative Models*. 2024. arXiv: 2401.07187 [stat.ML]. URL: https://arxiv.org/abs/2401.07187.

[8] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. New York, NY: Springer, 1995. DOI: 10.1007/978-1-4757-2440-0.