Rethinking Neural Networks with Benford's Law

Surya Kant Sahu*
The Learning Machines
surya.oju@pm.me

Abhinav Java
Delhi Technological University
java.abhinav99@gmail.com

Arshad Shaikh
BYJU's, India
arshaikh5775@gmail.com

Abstract

Benford's Law (BL) or the Significant Digit Law defines the probability distribution of the first digit of numerical values in a data sample. This Law is observed in many datasets. It can be seen as a measure of naturalness of a given distribution and finds its application in areas like anomaly and fraud detection. In this work, we address the following question: Is the distribution of the Neural Network parameters related to the network's generalization capability? To that end, we first define a metric, MLH (Model Enthalpy), that measures the closeness of a set of numbers to BL. Second, we use MLH as an alternative to Validation Accuracy for Early Stopping and provide experimental evidence that even if the optimal size of the validation set is known beforehand, the peak test accuracy attained is lower than early stopping with MLH i.e. not using a validation set at all.

1 Introduction and Related Works

Benford's Law (BL) has been observed in many naturally occurring populations, including the physical constants, populations of countries, areas of lakes, stock market indices, tax accounts, etc. [1]. Researchers have also discovered the presence of this law in natural sciences [2], image gradient magnitude [3], synthetic and natural images [4], etc. Attempts have been made to explain the underlying reason for BL's emergence for specific domains. However, a universally accepted explanation does not yet exist. The fact that BL occurs in many naturally occurring datasets, and the samples which don't not obey BL are probable anomalies, is one of the reasons why BL is also known as *"The Law of Anomalous Numbers"*. Due to this, BL has been used to ascertain fraud in taxing and accounting and for Anomaly Detection in the machine learning literature, such as detecting GAN-generated images [5].

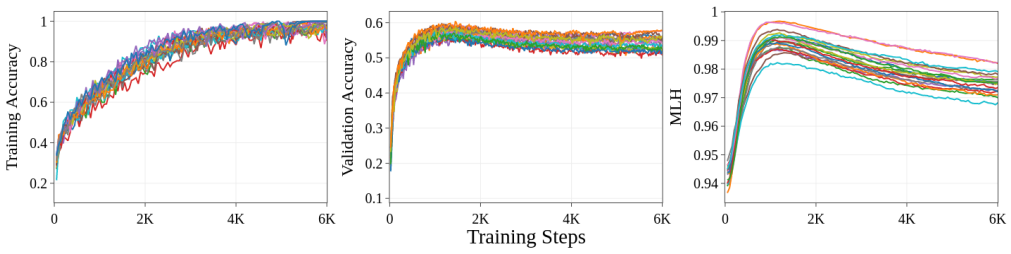

Figure 1: (Left to Right) Training accuracy, validation accuracy and MLH against training iterations. At around 1K iterations, the validation accuracy drops, while the training accuracy reaches 1.

*The Learning Machines is an independent student-driven research group. Github: https://github.com/The-Learning-Machines

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

1.1 Thermodynamics of Machine Learning

Previous work has established the formal connection of Thermodynamics and machine learning. [6] define four information-theoretic functionals, out of which, we focus on Relative Entropy S. It measures the entropy between the distribution $p(\theta|X,Y)$ that is assigned after training on data (X,Y) and prior $q(\theta)$ for model parameters θ. S can measure the risk of overfitting the parameters. The authors claim, this measure is intractable.

$$S \equiv \log \frac{p(\theta|X,Y)}{q(\theta)} \tag{1}$$

1.2 Free-Energy Principle and Information Criteria

Free-Energy Principle [7] is a well-known principle that tries to explain the mechanism of learning and behaviour in living beings (referred to as "agents"). This principle states that agents take actions to sensory input, and its own internal state through an internal model of the world. This model is updated based on the outcome of the action. The learning objective, according to the Free-Energy Principle, is to minimize "surprise" in addition to minimizing the complexity of the learned model.

Information-based Criteria are used frequently for model selection in the ML Community. Bayesian Information Criterion (BIC) and Akaike Information Criteria (AIC) [8] are some of the widely-used criteria for model selection.

$$AIC(m) = -2\log L(m) + 2p(m) \tag{2}$$

$$BIC(m) = -2\log L(m) + 2p(m)\log n \tag{3}$$

where m is a model, $L(m)$ is the error of the model m, $p(m)$ is the number of parameters of m, and n is the number of data-points used to learn m. Here, the number of parameters is used as the measure of model complexity.

In this work, we show that the metric we propose, MLH, could be a measure of model complexity, and hence use Free-Energy Principle for Early stopping using MLH.

2 Model Enthalpy

BL defines a probability distribution of a given sample's significant (leftmost) digit. The leftmost non-zero digit's occurrence in the observations of a population is log-uniform for several datasets, with 1 occurring the maximum number of times, followed by 2, 3, till 9. According to Benford's Law (BL) [9], the probability for a sample having a significant digit d is given as follows:

$$P_B = P(d) = \log_{10}(1 + \frac{1}{d}), d = 1, 2, 3, ..., 9 \tag{4}$$

We hypothesize that Neural Network weights might follow BL, similar to the pixels of RGB images. To that end, we devise a simple metric to measure the similarity of histograms of significant digits of model parameters. We define a metric called Model Enthalpy (MLH), that measures the correlation between Benford's Law and histogram of significant digits of a given set. MLH is based on the Pearson's Correlation Coefficient [10] is defined as follows:

$$MLH(\theta) = PearsonR(BinCount(\theta), P_B) \tag{5}$$

$$BinCount(\theta) = \frac{[f_0, f_1, ..., f_9]}{D_\theta} \tag{6}$$

Here, $BinCount(\theta)$ is the distribution of Significant Digits of network parameter set θ. P_B is the distribution defined by BL, f_k is is the frequency of significant digit k occurring in θ, D_θ is the dimensionality of θ. We did not include parameters that are initialized with a constant value, such as Bias and BatchNorm parameters. In our implementation, we multiply all elements in the set by a constant 10^{10} so that the resultant elements are greater than zero, and then take the first non-zero digit. This representation is required for a fast vectorized implementation of $BinCount(.)$. Note that

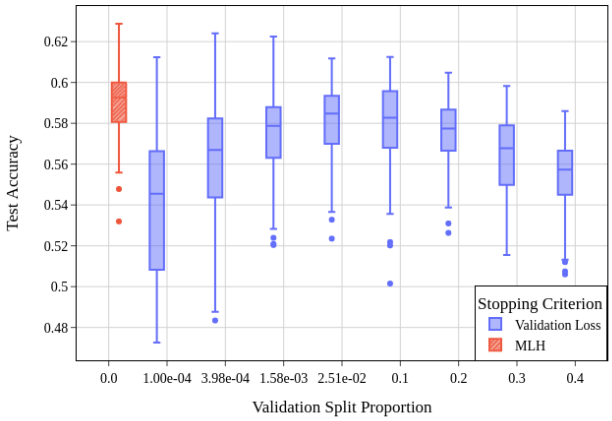

Figure 2: Comparison of Test Accuracies using (Red) where MLH is used as an Early Stopping criterion and (Blue) where validation proportions are used to dictate Early Stopping. We can observe that there is a noticeable difference in test accuracy for the methods and validation proportions used for early stopping. This indicates that MLH contains information about generalization capability of this model.

multiplying with a constant scalar doesn't change the distribution of significant digits due to BL's property of *Scale Invariance* [11].

In Fig. 1, we train over 100 shallow AlexNet-like models on CIFAR10 and track metrics during training iterations. It can be clearly seen that the proposed metric, MLH traces a similar trajectory as the Validation Accuracy.

3 Early Stopping with MLH

In this section, we present a direct application of MLH. In the previous section, we established that MLH is strongly correlated to validation accuracy. We use this result to replace validation set-based criteria for Early Stopping, while using the data saved as additional training data.

In Early Stopping, the stopping criterion is monitored throughout the training procedure, and if a certain predefined condition involving the criterion is met, the training is stopped. Usually, the criterion used is the accuracy on the validation set. Early Stopping based on validation set criteria require a validation set to be split off from the training data, which, depending on the size, results in a significant reduction in the amount of available training data. The size of the validation set can also be seen as a hyperparameter. Larger validation sets can result in poorer models due to lower amounts of training data. On the other hand, smaller validation sets can result in inaccurate estimates of generalization performance, and the criteria being unreliable, leading to premature or late stopping. The optimal size of the validation set finds the best trade-off, as observed in Fig. 2. But finding this optimal size of the validation set is non-trivial and requires multiple training runs.

For the experiments in this section, we use the publicly available CIFAR10 dataset, and a smaller AlexNet-like model without dropout to make sure the models overfit and hence make our observations concrete. We do a sweep of various validation set sizes and use *Validation Accuracy* as the Early Stopping metric. For each setting, we train 100 such models for computing confidence intervals[2]. Fig. 2 (Blue) illustrates the validation set size trade-off. For one set of models, Fig. 2 (Red), we use MLH as the Early Stopping criterion, and include validation data for training. Fig. 2 shows that even if the practitioner knows the optimal validation set size beforehand, the mean test accuracy is noticeably lower than when not using a validation set at all.

With the evidence presented in this section, we believe that MLH is strongly related to Eq. 1, because S measures the log-ratio of distribution over parameters, p and prior q. In the case of MLH, we

[2]The experiments can be reproduced with one day of compute on an RTX 3060 GPU.

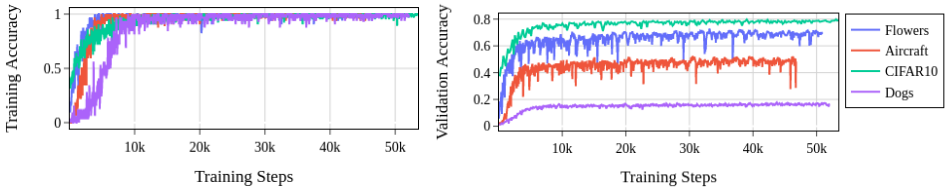

Figure 3: We train DenseNet121 with 7M parameters on four datasets. (Left) Training Accuracy of the architecture v/s. Training steps. Note that Training Accuracy reaches to 1. (Right) Validation Accuracy v/s. Training steps. We note that for the duration of 200 epochs, the model's Validation Accuracy doesn't fall significantly, if at all, contrary to AlexNet-like model in Fig. 1.

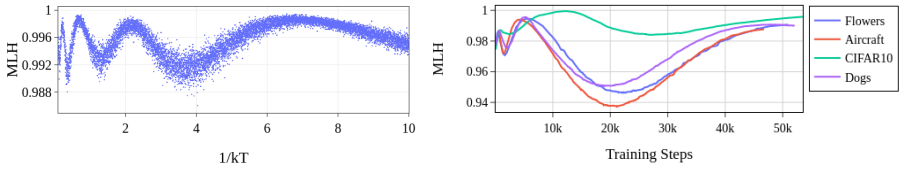

Figure 4: (Left) MLH of Energy states at different values of Temperature T. (Right) MLH of DenseNet121 weights on multiple datasets. We expected a different behaviour from Fig. 1, since there was no overfitting observed in Fig. 3.

assume the prior over parameters to be Benford's Law, i.e $q \cong P_B$. This would suggest that S is approximated by MLH. We hope to rigorously investigate this in the future. Assuming MLH is indeed a measure of model complexity, in the Early Stopping experiment, we were maximizing the Training Accuracy and MLH (higher MLH means lower overfitting), and hence the connection to Free-Energy Principle [7].

4 MLH and Deep Neural Networks

In the previous sections, we used shallow Alexnet-Like networks that were prone to overfitting. Recent work has shown that larger and deeper neural network architectures are robust to overfitting [12]. As observed in Figures 3 & 4 (Right), when we swap out the smaller AlexNet-like model with a larger DenseNet-121, we observe that the model doesn't overfit, i.e. the training loss reaches zero, however validation accuracy doesn't drop, instead plateaus, contrary to the shallow network. Due to this, we expect the behaviour of MLH to change. In Fig. 4, MLH oscillates for a deeper DenseNet121 architecture; we note this observation on multiple datasets. We believe that this distinct behaviour of MLH on shallow and deeper models (that don't overfit) support our hypothesis that MLH might be an indicator of generalization. However, MLH can not be used for Early Stopping for deeper networks in the way we in Sec. 3. We hope to explore deeper, state-of-the-art models in the future.

We connect this oscillatory pattern of MLH to a contribution by [1] where they find that for systems following Boltzmann-Gibbs statistics, such as an ideal gas in a sealed chamber, the mantissa distribution of energy states of particles oscillates around BL with change in temperature. This is illustrated by Fig. 4 (Left). Here, we run a simulation where we sample a large number of Energy states at a Temperature T with the probability density function for an energy state E from [1],

$$f(E) = \frac{1}{kT} e^{-\frac{E}{kT}} \tag{7}$$

Here, k is the Boltzmann Constant. We compute MLH of energies at $1/kT = 0.1$ to $1/kT = 10$ at 10000 equally-spaced values. Fig. 4 (Left) shows how MLH changes as a function of temperature T which strikingly resembles Fig. 4 (Right), where we plot 4 models trained on 4 different datasets [13, 14, 15, 16] and compute MLH of their weights.

Throughout this work, we assumed that MLH is a measure of model complexity, however, explaining why MLH contains this information would possibly also require us to answer why BL even emerges in the first place, which has remained unexplained since the phenomenon was discovered nearly two centuries ago.

5 Conclusion and Impact Statement

The research question of this work was: "Is Benford's Law related to Generalization in Neural Networks?". We first make the observation that Benford's Law is observed by weights of Neural Networks, and then study how closeness to BL changes during training. We show with strong evidence that BL is associated with generalization performance of a shallow Neural Network and hence, we use it as a replacement to validation metrics, eliminating the need of a validation set. We then study MLH on deeper networks which are robust to overfitting, where-in we observe a very different behaviour of MLH, which oscillates. To the best of our knowledge, this is the first work that shows the important connection of Benford's Law and Neural Networks, which can have potential applications, and may lead to a better understanding of nature of learning in Neural Networks and its possible connection to thermodynamics.

6 Acknowledgements

We sincerely thank Yannic Kilcher[3] for guiding and mentoring us on formulating experiments and writing the paper despite his very busy schedule.

References

[1] Lijing Shao and Bo-Qiang Ma. The significant digit law in statistical physics. *Physica A: Statistical Mechanics and its Applications*, 389:3109–3116, 05 2010.

[2] Malcolm Sambridge, Hrvoje Tkalčić, and A Jackson. Benford's law in the natural sciences. *Geophysical research letters*, 37(22), 2010.

[3] Jean-Michel Jolion. Images and benford's law. *Journal of Mathematical Imaging and Vision*, 14(1):73–81, 2001.

[4] E Acebo and Mateu Sbert. Benford's law for natural and synthetic images. In *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 169–176, 2005.

[5] Nicolò Bonettini, Paolo Bestagini, Simone Milani, and Stefano Tubaro. On the use of benford's law to detect gan-generated images. *arXiv preprint arXiv:2004.07682*, 2020.

[6] Alexander A. Alemi and Ian Fischer. Therml: Thermodynamics of machine learning, 2018.

[7] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[8] K. Burnham and David R. Anderson. Model selection and multimodel inference : a practical information-theoretic approach. *Journal of Wildlife Management*, 67:655, 2003.

[9] Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

[10] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

[11] Theodore P Hill. Base-invariance implies benford's law. *Proceedings of the American Mathematical Society*, 123(3):887–895, 1995.

[12] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.

[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[14] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[3] Youtube Channel: `https://www.youtube.com/c/YannicKilcher`

[15] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[16] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Some of our observations are limited to shallow neural networks.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is available at `https://github.com/The-Learning-Machines/RethinkingNNsWithBL`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Fig. 2
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]