What it means by learning in a neural network: easing the knot

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

What constitutes learning in humans and machines remains puzzling despite the unprecedented growth witnessed in both. Starting with a Perceptron and, in subsequent interrogation of multilayer perceptron (MLP) and deep linear neural network (DLN), this paper revisits and searches for new learning signatures and dynamics in artificial neural networks (ANN). Precisely, we consider the transport of the initial weight distribution to its final form while optimally balancing entropy (randomness in the weights) and statistical complexity, which captures the neural network's information storage structure. As found, training neural networks guided through complexity-entropy improves its reproducibility. In continuation, we further assess depth dependence and information flow using entropy-difference, KL-divergence of weight distribution between successive layers, and mutual information between input and hidden layers. Insights obtained so far in our ongoing analysis of perceptron learning are of immense importance, with applications ranging from explainable AI to understanding brain function.

1. Introduction

The Perceptron Minsky and Papert [9], Rosenblatt [12], an artificial mimic of the biological neuron, drawing its existence and theoretical underpinnings from how neurons enable learning in the brain Block [2], sparked the AI era. Learning in a perceptron-based network involves adjusting the weight distribution, eventually mimicking the transport of an initial weight distribution to its final form, through data-intensive training. Although a biological neuron and a Perceptron function possess considerable dissimilarities in working principle Hasson et al. [4], the gradients remain inseparable from the plasticity of biological neurons and the optimization sought in gradient descent in a perceptron-based network. In addition, a trained NN must also generalize well Allen-Zhu et al. [1], Jacot et al. [6], be reproducible, and frugal in training as well as inference, which are traits that must coexist in neural connectivity. It is important to note that, the strength and structure of neural connectivity get tuned by the distributional shift of weight strength that spells learning. Upon careful observation it is evident that, many counterintuitive objectives align, apparently inexplicably, defining the persistent black-box nature of deep learning models Rudin [13], Shwartz-Ziv and Tishby [15]. Sounds a bit overstated at its current form, but easing the learning riddle in neural networks has been the focus of this work, from a single perceptron to an MLP, using information-theoretic concepts Shwartz-Ziv and Tishby [15], Tishby and Zaslavsky [16], Tishby et al. [17], metrics, and statistical complexity Feldman and Crutchfield [3], Lopez-Ruiz et al. [8] as the lens.

Earlier works treated weight distribution tuning as an optimal transport problem. But how the structural correlation between components within a model shapes the learning, or whether it contributes to NNs' performance, in terms of generalizability, reproducibility, or computational effi-

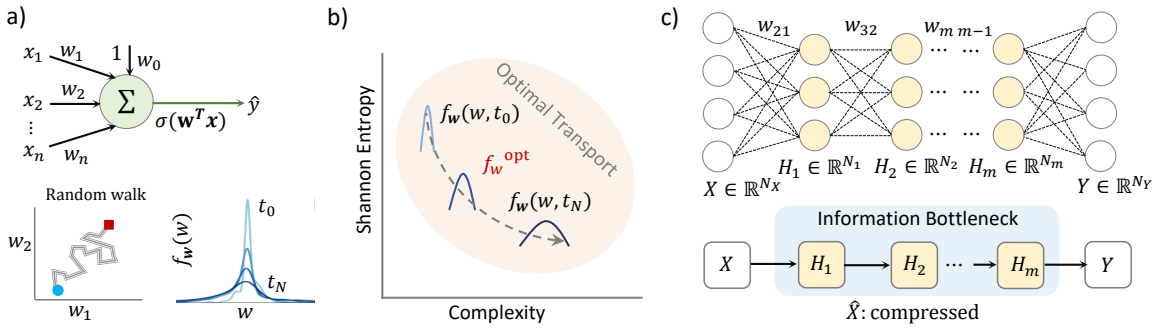

Figure 1: a) n-input perceptron (upper panel). An example random walk in the weight space $w_i \in \mathbb{R}^2$ (lower panel), and weight distribution $f_{\mathbf{w}}(w)$ transport as the training progresses. b) The proposed 2D Complexity-Entropy (C-E) plane to oversee the optimal transport $f_{\mathbf{w}}(w)$ in learning. c) A sample MLP, amenable to deep linear network (DLN) as well, with notion of mutual information $I(X; H_i)$ flow as in Information Bottleneck (IB) method.

ciency, remains puzzling. Additionally, the weight distribution in neural networks remains inexplicably different from the experimentally observed log-normal distribution of weights of synaptic plasticity in the brain. This leaves fundamental questions open and one such question is how the geometry of the plasticity, as frequently used in NNs, aligns with the brain. A closer look highlights a potential mismatch between the weight distribution geometry and the widely used gradient descent. Such a discrepancy is among many discrepancies that shape the learning in NNs. Moreover, approaches that rely on MDL Rissanen [11], such as soft weight sharing Nowlan and Hinton [10] to minimize weight MDL Hinton and Van Camp [5], reduce network complexity, improve model generalization, and help control overfitting. This approach also, like the others, does not explain learning in NN models and whether it is possible to analytically guide learning by structurally correlating the model components that shape the features of the governing probability distribution.

Recent evidence from neuroscience experiments suggests that brain connectomes are largely function-oriented Koulakov et al. [7], Scheffer et al. [14], attesting to the advantageous role of structural connectivity. For instance, the log-normal distribution of synaptic weights provides computational benefits by imposing a structural constraint. Motivated by the insights, we apply statistical Complexity, $C(\mathbf{w})$, to assess the structural correlation between system components and to interrogate the attainment of the final weight distribution on a 2D plane comprising complexity and entropy. Precisely, a fully connected network, as ideally represented by a perfect crystal in statistical mechanics, is in a minimum equilibrium Feldman and Crutchfield [3], Lopez-Ruiz et al. [8]. In contrast, a highly sparse network (e.g., an ideal gas) yields a high level of complexity that serves as an analog of equilibrium. In this work, we explore learning dynamics, myths, and beyond. Specifically, it interrogates a single-layer perceptron (SLP) and a multilayer perceptron (MLP) in search of the meaning of learning, that balances between order and disorder in a neural network Lopez-Ruiz et al. [8], Wiedermann et al. [18]. Precisely, this continuing work seeks an optimal balance among the trifecta of order, information, and equilibrium in neural network learning. Specific outcomes so far obtained are:

- Perceptron learning reflects the minimization of MSD
- Training error reduction slows down when the system is working away from the C($\mathbf{w}$)-H($\mathbf{w}$) extrema, suggesting that early stopping may relate to optimality between C($\mathbf{w}$) and H($\mathbf{w}$).
- Weight distribution satisfying the optimality of C($\mathbf{w}$)-H($\mathbf{w}$) achieves higher reproducibility maintaining a competitive classification accuracy.

2. Model and Methods

2.1. Assessment of Perceptron weight displacement

The MSD formulates as the ensemble average of the square of the displacement vector: $\text{MSD}(e) = \langle |\boldsymbol{w}(x_j + 1) - \boldsymbol{w}(x_j)|^2, \forall x_j \in \mathbf{x} \rangle$. To test if the weight displacement follows a pure random walk, we used the augmented Dickey-Fuller (ADF) test on the weight displacement on $\mathbb{R}^n$ space.

2.2. Statistical Complexity

Statistical complexity, denoted as $C(W)$, finds an analog in statistical mechanics where the quantity assesses how far a system is from the equilibrium and the underlying randomness in components' connectivity. In measure, C($\mathbf{w}$) is low in a system that is close to equilibrium, like an ideal gas, and holds a high value in disequilibrium, as is the case for crystals. The ideal-gas and crystal are analogous to fully-connected, highly-sparse networks, providing an avenue for navigating an optimal balance between the structure and information content of a system. The statistical complexity, $C(\mathbf{w})$, of a random variable $\mathbf{w}$ captures both the information stored in the system and the system's disequilibrium. In the context of NN, the weight distribution $\mathbf{w}$ evolves as training proceeds, thereby affecting both the Shannon entropy $H(\mathbf{w})$ and the disequilibrium, which measures the distance of $\mathbf{W}$ from its uniform distribution. So, learning by a NN is not only about accounting for the entropy of $\mathbf{w}$, but also about an increasing $C(\mathbf{w})$ indicating how efficient and accurate the NN is at organizing the information. Statistical complexity is defined as the product of KL divergence and JSD as follows:

$$
\begin{aligned}
&C(\mathbf{w}) = H(\mathbf{w}) \times JSD(\mathbf{w}||\text{Uniform}(\mathbf{w})), \text{Assume} f_{\mathbf{w}}(w) = P, \text{and Uniform}(\mathbf{w}) = Q \\
&\text{JSD}(P||Q) = \frac{1}{2} D_{KL}(P||R) + \frac{1}{2} D_{KL}(Q||R), \text{ where } R = \frac{P+Q}{2} \\
&H(P) = \sum_{x \in \mathbf{X}} p(x) \log \frac{1}{p(x)}, \text{and between distribution A, B } D_{\text{KL}}(A||B) = \sum_i A(i) \log \frac{A(i)}{B(i)}
\end{aligned} \tag{1}
$$

2.3. Mutual Information in Hidden Layers

We estimate the mutual information (MI) between the input X and successive hidden layers, T, and the mutual information between output Y and hidden layers T, for layer sizes 2, 6, and 16. The information path follows the relationships as follows:

$$
I(X;T_1) \geq I(X;T_2) \geq I(X,T_3) \geq \cdots \geq I(X;T_k) \tag{2}
$$

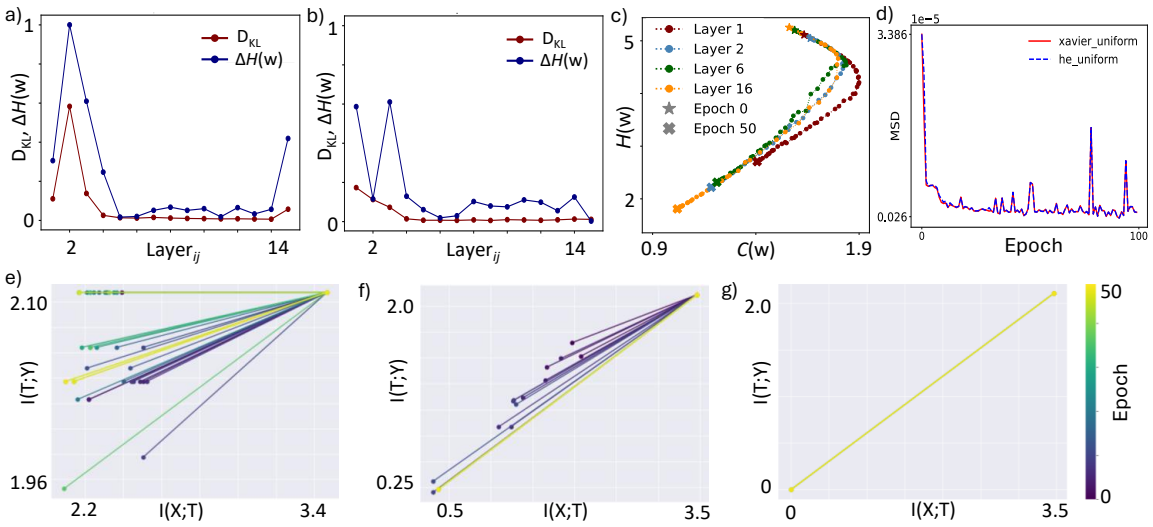

Figure 2: Comparison of Entropy ($\Delta H(\mathbf{w})$ and $D_{KL}(\mathbf{w}||\mathbf{w}_{i+1})$: a) For MNIST trained on 16-layer MLP, b) Same configuration but for CIFAR-10. c) Beyond Layer 6, final epoch drives the system to non-optimal region of C-E plane. d) MSD of perceptron indicates a reduction as training epoch increases. e-g) Mutual information between input, output and successive layers.

As shown in the information plane $I(T;Y)$ against $I(X;T)$ (Fig. 2), increasing the number of layers ceases to mutual information across epochs. Mutual information between X and layers closer to X is greater than MI for layers farther away.

3. Results

3.1. Entropy flow and KL-divergence across layer are qualitatively similar to error reduction

Entropy difference, between a layer L_i to its immediate next L_{i+1}, capture the difference in information between the layers. The difference, as shown in Fig. 2(a-b), gradually approaches to its minimized level as the network goes deeper reassuring that deepening a network doesn't introduce much additional information, and appears qualitatively similar for both MNIST (Fig. 2a) and CIFAR-10 (Fig. 2b) datasets. Specifically, around layer 6, both $(\Delta H(\mathbf{w})$ and $D_{KL}(\mathbf{w}||\mathbf{w}_{i+1})$ become sufficiently low. To further analyze, we project per-layer training response over all epochs on a Complexity-Entropy (C-E) plane that shows a gradual progression towards complete order and high disequilibrium state, beyond which model progresses to sub-optimal region on the C-E plane, intriguing the need of a thorough analysis of training and learning dynamics on the C-E plane. For successive epochs, the shift in mutual information between earlier layers and Y, and deeper layers and Y, increases gradually (Fig. 2f). The transformation in mutual information as number of layers are increased, may be a reflection of the vanishing gradient issues of large network of MLP, and remains a part of our ongoing investigation.

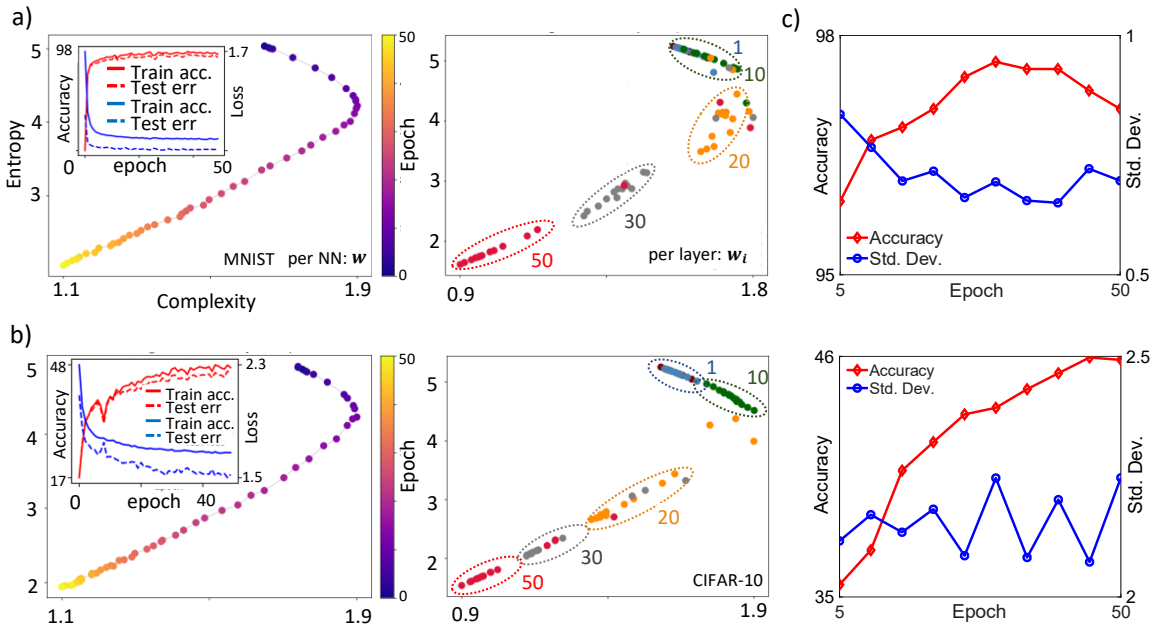

Figure 3: Model performance on Complexity-Entropy plane: a) considers all layer weights $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_{16}]$ to form a weight distribution for $C(\mathbf{w})$ and entropy $H(\mathbf{w})$ calculation. b) Here, per layer (i) weight distribution, $p(\mathbf{w}_i)$, is used to calculate $C(\mathbf{w}_i)$ and $H(\mathbf{w}_1)$, and each circle represents C-E projection of all layers. a-b) For MNIST (upper panel) and CIFAR-10 (lower panel), early training epochs increase complexity, but this effect is reversed after about 10 epochs. Training error, shown in the inset plot, falls sharply by 10 epochs and does not decrease further thereafter. This trend remains similar for both datasets tried. c) Mean accuracy and standard deviation over 20 testing results using weights $\mathbf{W}$ obtained after 5, 10, 20, 30, 50 training epochs, both for MNIST (upper) and CIFAR-10 (lower). A low stdev indicates better reproducibility — MLP achieves improved reproducibility and competitive accuracy when trained in the optimal C-E region.

3.2. Weights optimal on C-E plane achieves enhanced reproducibility

The C-E plane posits training on order and disequilibrium plane, where the optimality lies away from extrema of both Feldman and Crutchfield [3], Lopez-Ruiz et al. [8]. For instance, when MLP is trained for 50 epochs, the weight tuning drives the model towards a highly ordered (low information) and low disequilibrium region, compromising both the network complexity and the information stored in it (see Fig. 3). Interestingly, complexity increases in early epochs and reverses its progression at a point that resembles the drastic reduction in training error (see inset of Fig. 3(a-b), left panel). Interestingly, such trend of model complexity remains similar regardless of layer-wise (Fig. 3(a-b), right panel) or per-NN analysis (Fig. 3(a-b), left panel), as well as over alternative datasets (MNIST, Fig. 3a and CIFAR, Fig. 3b). An optimal region, as theorized in Lopez-Ruiz et al. [8], resides near projections of training dynamics of epoch 35 to 20, strengthening both complexity and entropy on the C-E plane. To assess how the system performs for a C-E-guided weight distribution tuning, we calculated 20 accuracies, each performed 20 times on randomly chosen 500

test samples, and reported the mean accuracy and standard deviation. As observed, the weight distribution obtained for epochs 20 to 35 has a smaller standard deviation while maintaining competitive accuracy (both on MNIST, Fig. 3c, upper and CIFAR-10, Fig. 3c, lower), suggesting that better reproducibility may relate to the optimal region identified in the complexity-entropy plane.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[2] Hans-Dieter Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

[3] David P Feldman and James P Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238(4-5):244–252, 1998.

[4] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[5] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

[6] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[7] Alexei A Koulakov, Tomáš Hromádka, and Anthony M Zador. Correlated connectivity and the distribution of firing rates in the neocortex. *Journal of Neuroscience*, 29(12):3685–3694, 2009.

[8] Ricardo Lopez-Ruiz, Hector L Mancini, and Xavier Calbet. A statistical measure of complexity. *Physics letters A*, 209(5-6):321–326, 1995.

[9] Marvin Minsky and Seymour A Papert. *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. MIT press, 2017.

[10] Steven J Nowlan and Geoffrey E Hinton. Simplifying neural networks by soft weight sharing. In *The mathematics of generalization*, pages 373–394. CRC Press, 2018.

[11] Jorma Rissanen. Stochastic complexity and modeling. *The annals of statistics*, pages 1080–1100, 1986.

[12] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[13] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[14] Louis K Scheffer, C Shan Xu, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth J Hayworth, Gary B Huang, Kazunori Shinomiya, Jeremy Maitlin-Shepard, Stuart Berg, et al. A connectome and analysis of the adult drosophila central brain. *elife*, 9:e57443, 2020.

[15] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[16] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

[17] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[18] Marc Wiedermann, Jonathan F Donges, Jürgen Kurths, and Reik V Donner. Mapping and discrimination of networks in the complexity-entropy plane. *Physical Review E*, 96(4):042304, 2017.

Appendix A. Experimental Setup

The MLP consists of an input layer, 16 hidden layers, and an output layer. The input layer has 784 neurons for MNIST and 3072 neurons for CIFAR-10, each hidden layer has 512 neurons, and the output layer has 10 neurons for classification. We used ReLU as activation between the hidden layers, and PyTorch's CrossEntropyLoss applies softmax to the output layer internally. Each hidden layer also includes batch normalization and a dropout rate of 0.3. The datasets used are MNIST (28×28 grayscale) and CIFAR-10 (32×32 RGB), both containing 10 classes. MNIST has 60,000 training samples, and CIFAR-10 has 50,000, and both provide a test set of 10,000 samples. For training, we used 40,000 samples per epoch, sampled from the training set (a different subset in each epoch). We evaluated the model on the full 10,000 test samples after each epoch. We trained for 50 epochs with a batch size of 32, the Adam optimizer with a learning rate of 0.001 and a weight decay of 1e-4. We pursued two initialization strategies: He-uniform and He-normal.