Rethinking Activation Functions in Deep Learning: A Theoretical Physics Perspective

Alejandro Chinea Manrique de Lara

Departamento de Física Fundamental UNED (Spain)

Abstract

Recent studies have established a strong connection between the mathematical properties of nonlinear activation functions employed in the fundamental computational unit of deep learning machines- the artificial neuron- and the emergence of dynamical isometry. The latter denotes a condition on the singular value spectrum of the input-output Jacobian of the network mapping, whereby gradients propagate stably across depth (i.e., neither exploding nor vanishing during training), thus promoting faster convergence and improved training stability. However it remains unclear whether general theoretical principles exists that govern the performance of these models more broadly as a function of the activation functions employed. In this work, the two most widely used families of nonlinear activation functions-- sigmoidal and ReLU-- are analyzed from a thermodynamics perspective. It is shown that ReLU functions exhibit a thermodynamic signature that is fundamentally distinct from that of sigmoidal functions, leading to qualitatively different thermodynamic behavior in the resulting networks. This distinction impacts not only learning dynamics and generalization performance but also the thermodynamic efficiency of these models. These theoretical results provide new insights into the foundations of deep learning. Their implications are further supported by a series of benchmark experiments specifically designed to empirically assess the validity of the proposed framework.

Introduction

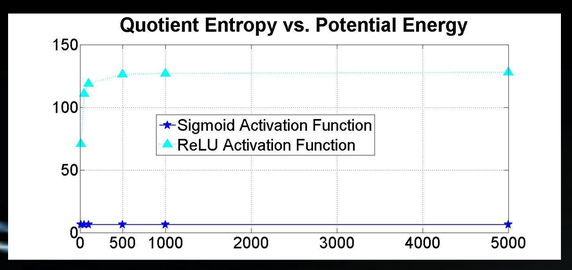
The rectified linear unit (hereafter ReLU) is the most widely used activation function in deep learning. Its success stems from its high computational efficiency and its ability to mitigate the vanishing gradient problem [1]. This speedup is vital for large-scale models utilizing hardware accelerators (e.g., GPUs) as well as for mobile deployment. Ultimately, efficient gradient propagation in ReLU networks yields faster convergence, enabling the training of exceptionally deep models [3][5][6]. Using the methodology from [2], this paper analyzes shallow ReLU and sigmoidal networks from a thermodynamic perspective to evaluate their generalization performance and the computational costs of their learning and recall phases. It is shown that ReLU networks yield better average generalization because they store less energy while maintaining higher entropy. Surprisingly, and contrary to popular belief, ReLU networks exhibit worse thermodynamic efficiency than sigmoid and hyperbolic tangent architectures. Consequently, depending on the unit count and input dimensionality, the computational costs for training can actually be higher for ReLU networks due to how heavily the activation function dictates the system's thermodynamic regime.

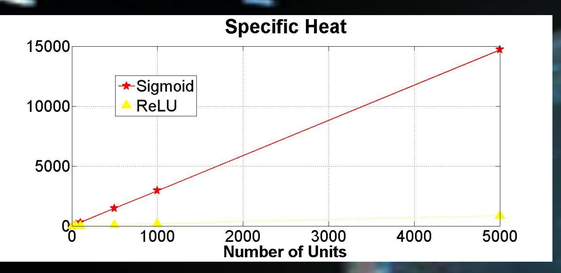
Theoretical Results

Entropy and internal energy are derived from the expression of the generating function of energies, with the stored potential energy calculated by subtracting the total neuron count from the internal energy. The mathematical expressions of the entropy (S) and potential energy (U ♮) are shown below, and also the graphs that display the evolution of the entropy-to-potential-energy ratio (left) and heat capacity as a function of input space dimension for 2xg_1x1 shallow architectures, respectively. The entropy-to-potential-energy ratio quantifies expected generalization performance and learning speed (epochs required to reach target error limits). The heat capacity governs the thermodynamic efficiency of information processing. Higher specific heat enhances the network's heat absorption capacity during computation, minimizing heat dissipation. On a Von Neumann computer architecture, this fact reduces computational runtimes during optimization and inference [2]. Theoretically, these metrics indicate that for a given structural complexity, shallow ReLU networks lead to a more stable thermodynamic regime than sigmoidal or hyperbolic tangent architectures. This regime provides increased information-storage capacity (higher entropy) and a smoother loss landscape geometry (lower potential energy), albeit at the expense of lower information-processing efficiency. Consequently, while shallow ReLU networks inherently learn and generalize better than sigmoidal networks, they scale with a distinct computational runtime penalty as the total number of neurons grows.

$$S = g_1 \log\left(\lambda^1_{m,p,g_0} + \sum_{\theta=\delta_1}^{g_0 mp} e^{-\beta\theta}\lambda^\theta_{m,p,g_0}\right) + \log\left(\lambda^1_{m,\theta_1,g_1} + e^{-\beta\Delta}\lambda^2_{m,\theta_1,g_1}\right) + \beta\Delta e^{-\beta\Delta}\frac{\lambda^2_{m,\theta_1,g_1}}{\lambda^1_{m,\theta_1,g_1} + e^{-\beta\Delta}\lambda^2_{m,\theta_1,g_1}} + \frac{\beta g_1 \sum_{\theta=\delta_1}^{g_0 mp}\theta e^{-\beta\theta}\lambda^2_{m,\theta_1,g_1}}{\lambda^1_{m,p,g_0} + \sum_{\theta=\delta_1}^{g_0 mp} e^{-\beta\theta}\lambda^\theta_{m,p,g_0}}$$

$$U_\phi = \Delta e^{-\beta\Delta}\frac{\lambda^2_{m,\theta_1,g_1}}{\lambda^1_{m,\theta_1,g_1} + e^{-\beta\Delta}\lambda^2_{m,\theta_1,g_1}} + \frac{g_1 \sum_{\theta=\delta_1}^{g_0 mp}\theta e^{-\beta\theta}\lambda^2_{m,\theta_1,g_1}}{\lambda^1_{m,p,g_0} + \sum_{\theta=\delta_1}^{g_0 mp} e^{-\beta\theta}\lambda^\theta_{m,p,g_0}}$$





Methods

Grounded in the statistical physics framework of [2], this model describes a shallow architecture with g_0 inputs, a single hidden layer of g_1 ReLU units, and a single-unit sigmoidal output layer, maintaining consistent notation throughout. This approach treats deep networks as thermodynamic systems of particles (neurons) subject to an external field (network topology) that restricts the system's degrees of freedom and alters its accessible phase space volume. Here, entropy quantifies architectural power, internal energy represents loss landscape complexity, and specific heat measures computational speed and thermodynamic efficiency. Accordingly, the generating function of energies and the coefficients that appears in the mathematical expression reads:

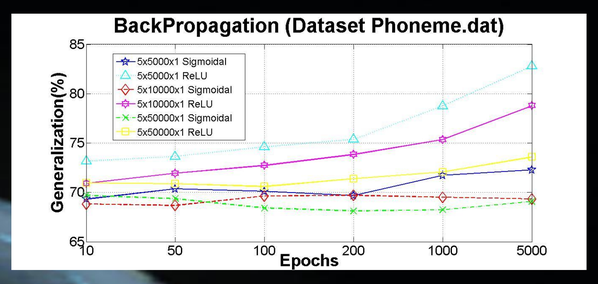
$$M(z) = \left(\lambda^1_{m,p,g_0} z^\mu + \sum_{\theta=\delta_1}^{g_0 mp}\lambda^\theta_{m,p,g_0} z^{\mu+\theta}\right)^{g_1}\left(\lambda^1_{m,\theta_1,g_1} z^\mu + \lambda^2_{m,\theta_1,g_1} z^{\mu+\Delta}\right)$$

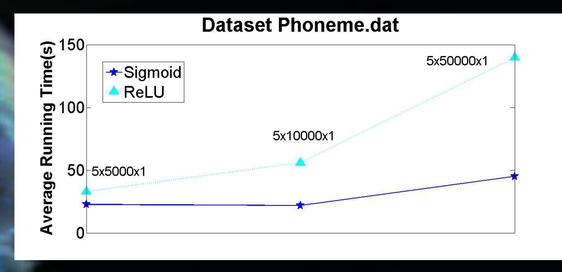
$$\lambda^1_{m,p,g_0} = [z^\mu]X(z)^{ReLU} = \sum_{n=g_0}^{\delta_1-1}\sum_{l=0}^{n}(-1)^l\binom{g_0}{l}\binom{g_0 - mpl - 1}{g_0 - 1}$$

$$\lambda^\theta_{m,p,g_0} = [z^{\mu+\theta}]X(z)^{ReLU} = \sum_{l=0}^{\theta}(-1)^l\binom{g_0}{l}\binom{\theta - mpl - 1}{g_0 - 1}$$

Discussion

To validate the preceding theoretical results, we evaluated performance using datasets from the European ELENA research project [7]. The graphs depict the performance metrics for a 5xNx1 architecture (N = 5,000, 10,000, 50,000) trained on phoneme.dat via backpropagation. The top graph outlines generalization, while the bottom graph displays average computational runtimes calculated across 1,000 independent learning phases of 50 epochs each.The generalization curves confirm that shallow ReLU networks achieve superior generalization bounds and faster convergence speeds than sigmoidal networks. Conversely, the training-phase computational costs expose a critical trade-off. Contrary to conventional machine learning assumptions, the empirical runtimes validate our theoretical predictions: shallow ReLU networks exhibit worse thermodynamic efficiency than sigmoid and hyperbolic tangent networks, causing training runtimes to escalate alongside hidden layer width. This phenomenon was replicated on the eight-dimensional Gauss8D.dat benchmark. Crucially, this runtime penalty does not affect the recall (inference) phase, which remains consistently faster for ReLU units due to their minimal algorithmic complexity on standard hardware. For low-dimensional data, such as the two-dimensional Clouds.dat dataset, this thermodynamic inefficiency only manifests when the network scales to 10,000 units or more.





References

[1] Bishop C.M., (1995).Neural Networks for Pattern Recognition. Oxford University Press.
[2] Chinea A.(2023). On the Theory of Deep Learning: A Theoretical Physics Perspective (Part I). https://doi.org/10.1016/j.physa.2023.129308
[3] Dubey, S.R., Singh S.K., and Chaudhuri B.B. (2022). Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. Neurocomputing, vol. 503, pp. 92-108.
[4] Efron B. (1979). Bootstrap Methods: Another Look at the Jacknife. The Annals of Statistics, Vol. 7, No. 1, pp. 1-26.
[5] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Pp. 315-323.
[6] Krizhevsky A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, Vol. 25, pp. 1097-1105.
[7] Blayo F., Cheneval Y., Guerin-Dugué A., et al. (1995), Enhanced Learning for Evolutive Neural Architecture ESPRIT Basic Research Project Number 6891, Deliverable R3-B4-P, Task B4 (Benchmarks).

Acknowledgments

This research was funded by the Nicolaus Copernicus University (UMK) under the DAMSI grant (Dynamics, Mathematical Analysis and Artificial Intelligence) of the University Center of Excellence and performed during a visiting researcher position of four months starting in September 2024 at the Department of Physics of the UMK.