Multiple Descents in Deep Learning as a Sequence of Order-Chaos Transitions

Wei Wenbo[a], **Chong Jia Le Nicholas**[a], **Lai Choy Heng**[a], **Feng Ling**[b]

[a] *Department of Physics, National University of Singapore, Singapore*

[a] *Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore*
fengl@ihpc.a-star.edu.sg

1. Introduction

Understanding the training dynamics of deep learning models is essential for improving their generalization, optimization, and robustness. The training process of deep neural networks involves navigating complex, high-dimensional parameter spaces, influenced by model complexity, dataset characteristics, and learning algorithms. Double descent challenges traditional bias-variance trade-offs and has been observed in neural networks, but many questions remain about its mechanisms and extensions to different architectures [1, 2].

In this study, we introduce a novel "multiple-descent" phenomenon in Long Short-Term Memory (LSTM) networks, where test loss cycles repeatedly, increasing slowly and then abruptly decreasing during training. We found out that that these cycles are linked to phase transitions between order and chaos in the network's dynamics, where the abrupt drop of test loss is often due to the sudden chaos order transition. Our research focuses on an LSTM trained on the Large Movie Review Dataset for sentiment analysis, revealing that the optimal performance occurs at the first transition from order to chaos, where the network's "edge of chaos" is widest, facilitating the best exploration of weight configurations.

2. Methods

We developed a basic LSTM model to perform sentiment analysis on the Large Movie Review Dataset, which contains 50,000 labeled movie reviews from IMDb [3]. The dataset was split into 70% for training and 30% for testing, with reviews padded or truncated to a fixed length. The model was intentionally over-trained to 1000 epochs to induce overfitting and observe dynamic behaviors. We used the Adam optimizer with a standard learning rate to train the model. To analyze the order and chaos phases, we conducted an asymptotic stability analysis similar to Ref [4], perturbing the LSTM's recurrent states and measuring how these perturbations propagate over time, providing insights into the network's stability and phase transitions.

3. Results on Multiple Descents and Order-Chaos Transitions

Our experiments, conducted over 1000 training epochs, uncovered two significant findings regarding the training dynamics of the LSTM model. The main results are illustrated in Figure 1. First, after the model began overfitting around epoch 500, we observed a striking pattern of multiple descents in the test loss. Specifically, the test loss exhibited eight distinct cycles over 500 epoches, with each cycle characterized by a gradual increase followed by a sudden, sharp decline within a single epoch. For instance, notable drops occurred near epochs 600, 700, 850, and 1,000. These cycles were closely mirrored by changes in the asymptotic distances measurements, a metric derived from our stability analysis. When the test loss increased, the asymptotic distance also rose, indicating a chaotic phase, but both metrics experienced abrupt drops at the end of each cycle, signaling a transition back to an ordered phase.

To further validate these phase transitions, we visualized the reduced sum of the LSTM's asympototic output vector ($\mathbf{h_{1599}} \cdot \mathbf{1}$) across 500 test samples per epoch. In ordered phases, the reduced sums converged to one or a few values, reflecting stability. In contrast, during chaotic phases, these sums scattered widely, indicating instability. This pattern confirmed that each descent in test loss corresponded to a transition from chaos to order.

Second, and perhaps more importantly, we found that the global optimal performance—achieved at epoch 114 with a test accuracy of 88.34%—coincided with the first transition from an order to a chaotic state. This transition was marked by a significant increase in asymptotic distance and a wide scattering of reduced sums of the asymptotic activations $\mathbf{h_{1599}}$ of each review sample, reflecting a broad "edge of chaos." This wide transition, occurring between epochs 50 and 100, resembled the slow initial order-to-chaos shift observed in non-linear dynamical systems, such as the one dimensional tanh map ($k_t = rk_{t-1}(1-\tanh(k_{t-1}))$). In contrast, subsequent transitions during the multiple descents after epoch 450 were narrower, with less pronounced bifurcations, suggesting faster transitions to chaos. These findings indicate that the first order-to-chaos transition provides the widest window for exploring optimal weight configurations, explaining why it yields the best model performance.

4. Discussion

Our discovery of multiple descents linked to order-chaos transitions in LSTM training offers a novel perspective on deep learning dynamics. Unlike the double descent phenomenon [2], which focuses on model complexity or training duration,

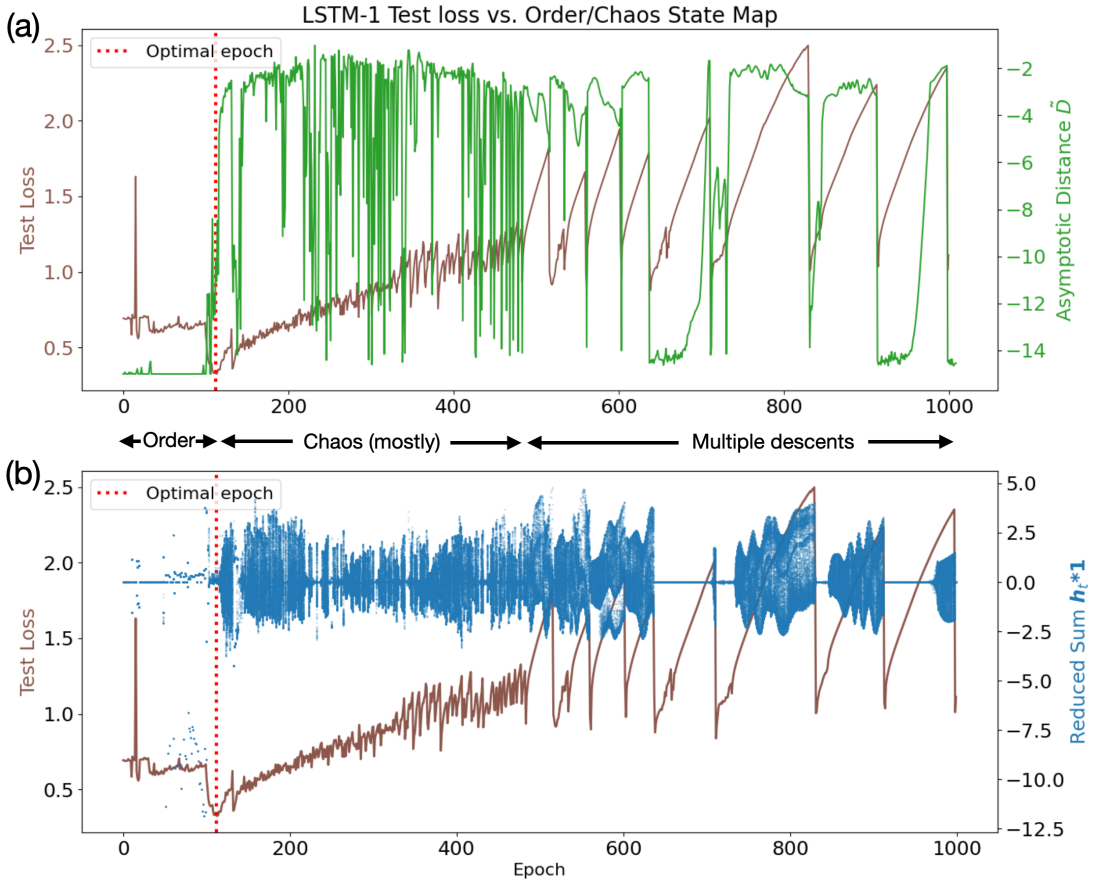

Fig. 1: Multiple descents through a sequence of order chaos transitions during the training process of LSTM. (a) The average asymptotic log distances $\tilde{D}$ (green) under perturbation is used to indicate order/chaos states. The optimal epoch of LSTM-1 is 114 with an accuracy of 88.34%. Multiple descents are seen in the overfitting regime at epochs > 450. When the asymptotic distance is at -15, it means two slightly different initial input values will converge to the same value at long enough iterations of the LSTM cell, indicating order phase. If the asymptotic distance is large, it means the model is at chaotic phase. (b) The 'bifurcation map' (blue) is shown together with the test loss (brown). The 'bifurcation map' is drawn by plotting the reduced sum $\boldsymbol{h}_{1599} \cdot \mathbf{1}$ for each of the 500 review samples at every epoch. Similarly, if the different samples converge to the same value, it indicates order phase. If the samples spread out, it indicates chaotic phase.

our findings highlight the temporal evolution of network stability during overfitting, with repeated cycles of chaos and order. The critical insight is that the global optimal performance occurs at the first order-to-chaos transition, where the "edge of chaos" is widest. This aligns with theories suggesting that networks perform best when balanced between order and chaos, optimizing learning capacity [4]. Our results draw parallels with non-linear dynamical systems like the tanh map, where the initial transition to chaos is slow and broad, enabling extensive exploration of parameter space. This suggests that the qualitative patterns we observed may generalize to other deep learning models, opening new avenues for research into training strategies. Practically, these insights could guide practitioners to identify optimal training epochs by monitoring phase transitions, potentially enhancing model generalization and robustness beyond traditional overfitting mitigation techniques.

References

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[2] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[3] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[4] Ling Feng, Lin Zhang, and Choy Heng Lai. Optimal machine intelligence at the edge of chaos. *arXiv preprint arXiv:1909.05176*, 2019.