0.2

THE UNIVERSITY of EDINBURGH

Neural Networks Thrive at the Edge of Chaos!

Rajit Rajpal

Problem

There is a huge gap between theory and practice in Neural networks. There is no widely accepted theoretical framework for generalization. SGD tends to be quite successful in optimizing Neural Network loss functions despite the high-dimensional non-convex structure. The geometry of loss landscapes [3] have also been studied that use the Hessian as a measure of curvature. Flat minima lead to better generalization while sharper minima lead to poor generalization. This is problematic as it depends on model parameterization. It has been proposed that it is the connectivity of the local minima that yield good generalization [5]. Connections have been drawn between the energy landscape of mean-field glasses and loss landscape of DNNs [1] implying that local minima can achieve good generalization while global minima may lead to overfitting.

Goal: To develop a theoretical framework to explain generalization in neural networks that is independent of model parameterization

Criticality

Criticality is a common occurrence in nature, as seen in the sandpile model where a system transitions from stable to unstable behavior as more sand grains are added. This evolution towards criticality makes the system highly sensitive to small perturbations, leading to cascading events and demonstrating its abrupt, scale-free responses [1].

Edge of Chaos Condition

The EoC condition for a nonlinear dynamical system $\mathbf{x_{t+1}} = \mathbf{f}(\mathbf{x_t})$:

$$\frac{1}{\sqrt{N}}\|\mathbf{J}^*\|_F = 1$$

where $\mathbf{J}^*$ is the Jacobian of $\mathbf{f}(\mathbf{x}^*)$ where $\mathbf{x}^*$ is the asymptotic value of $\mathbf{x}$.

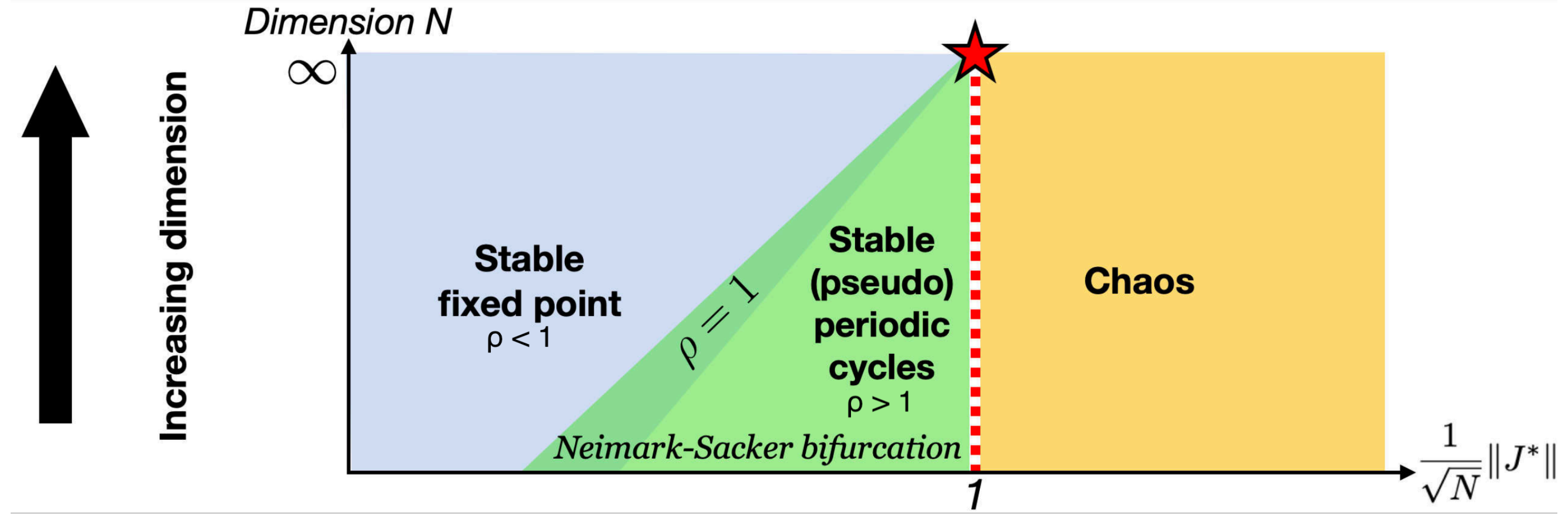
Increasing dimension
Dimension N
∞
Stable fixed point
ρ < 1
$\rho = 1$
Stable (pseudo) periodic cycles
ρ > 1
Chaos
Neimark-Sacker bifurcation
1
$\frac{1}{\sqrt{N}}\|J^*\|$

Fig 1: The three training phases of neural networks. Source: [2]

Connection with Neural Networks

Deep Neural Networks follow three phases of training [*Fig 1*] [6]:

1) stable fixed point phase
2) the periodic cycle phase
3) the chaotic phase

Fig 2 demonstrates that optimal generalization occurs at the boundary between periodic cycle and chaotic phase (Edge of Chaos [EoC]) [2].

<u>Limitation</u>: Input dimension = Output dimension to evaluate the asymptotic Jacobian $\mathbf{J}^*$.

<u>Note</u>: The asymptotic state is invariant to model reparameterization.

Empirical Results

Fig 2 shows that for a MLP, the lowest test loss is achieved where:

$$\frac{1}{\sqrt{N}}\overline{\|\mathbf{J}^*\|_F} \approx 1,$$

where $\frac{1}{\sqrt{N}}\overline{\|\mathbf{J}^*\|_F}$ is the geometric mean estimated after each epoch.

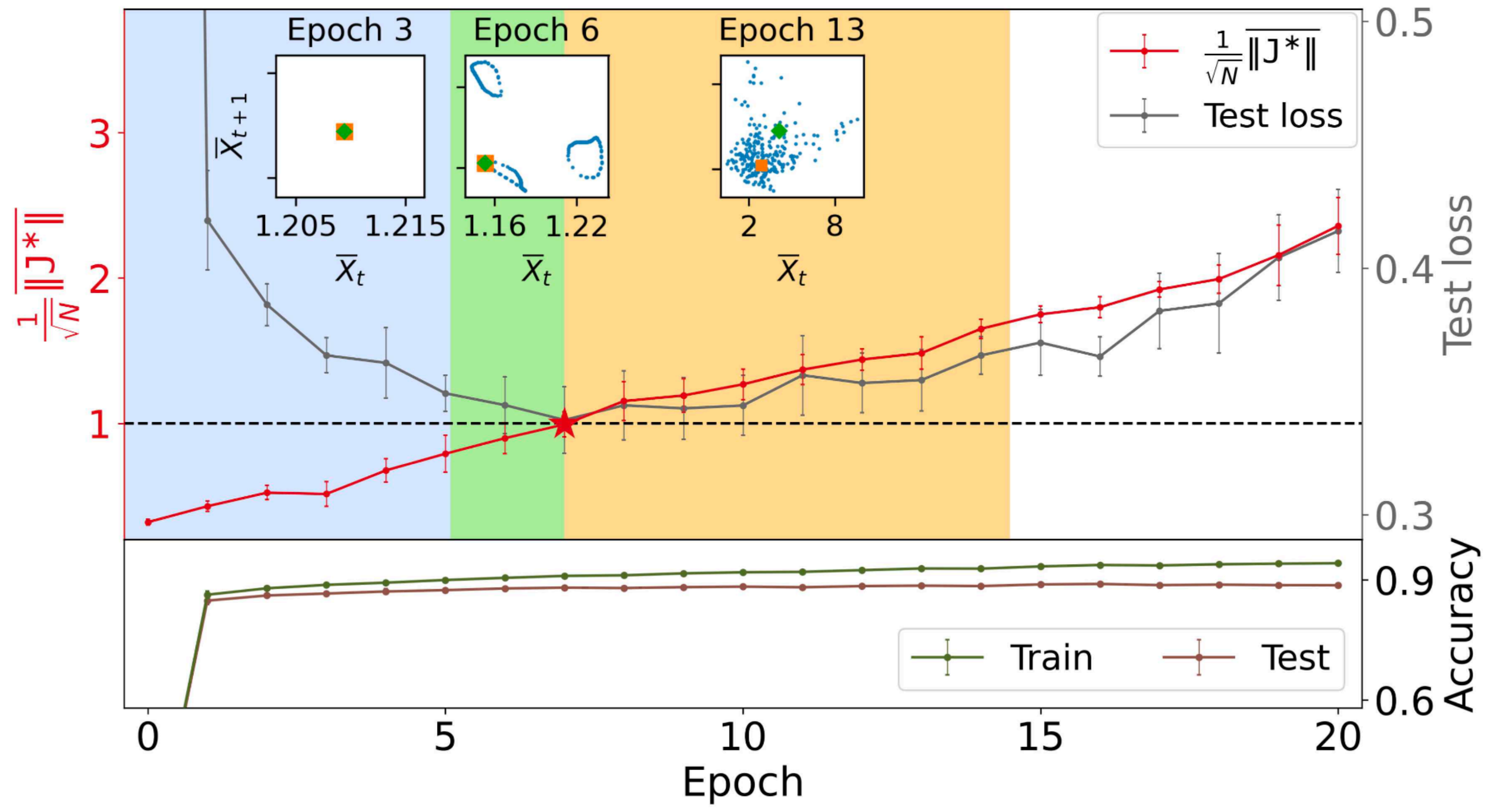

Fig 2: MLP trained on Fashion MNIST. Same color scheme as Fig 1. Source: [2]

Analytical Edge of Chaos Condition

The EoC condition for single-hidden layer neural networks with tanh activation function is:

$$J^2 \int D_z sech^4(J_0\mu + J\sqrt{q_0}z) = 1,$$

where $q_0 = \frac{1}{N}\Sigma\mu_i^2, D_z = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, and J_0, J^2 correspond to the mean and variance of the weights [2]. This is identical to the spin-glass boundary in the Sherrington-Kirkpatrick spin-glass model [5].

Chaotic effect of Regularization

Regularization is used in Neural Networks to improve generalization [3]. Weight decay is the most popular method. However, selecting the optimal strength of regularization is still an open problem.

<u>*Hypothesis*</u>: When various training parameters act together with the regularization strength to bring the model's steady state at the edge of chaos, its performance should be the most optimal [2].

<u>*Results*</u>: Weight decay is applied to the model. When $\lambda = 9 \times 10^{-5}$, the model ends right at the edge of chaos; meanwhile, its test accuracy and early stopping performance happen to be the best, indicating the optimal model generalization performance [*See Fig 3*].

Relation of training parameters in the ordered phase

In the ordered phase ($J^2 < 1$), J^2 will increase linearly with time:

$$J^2 = A \times Epoch + C,$$

Where A depends on training parameters and C depends on initialization. The slope A influences how fast a model will arrive at the edge of chaos ($J^2 = 1$) .

$$A = \frac{\eta}{(1-\alpha)B}.D$$

where η is the learning rate, α is the momentum of SGD, B is the batch size, and D is a constant which depends on network structure and training data.

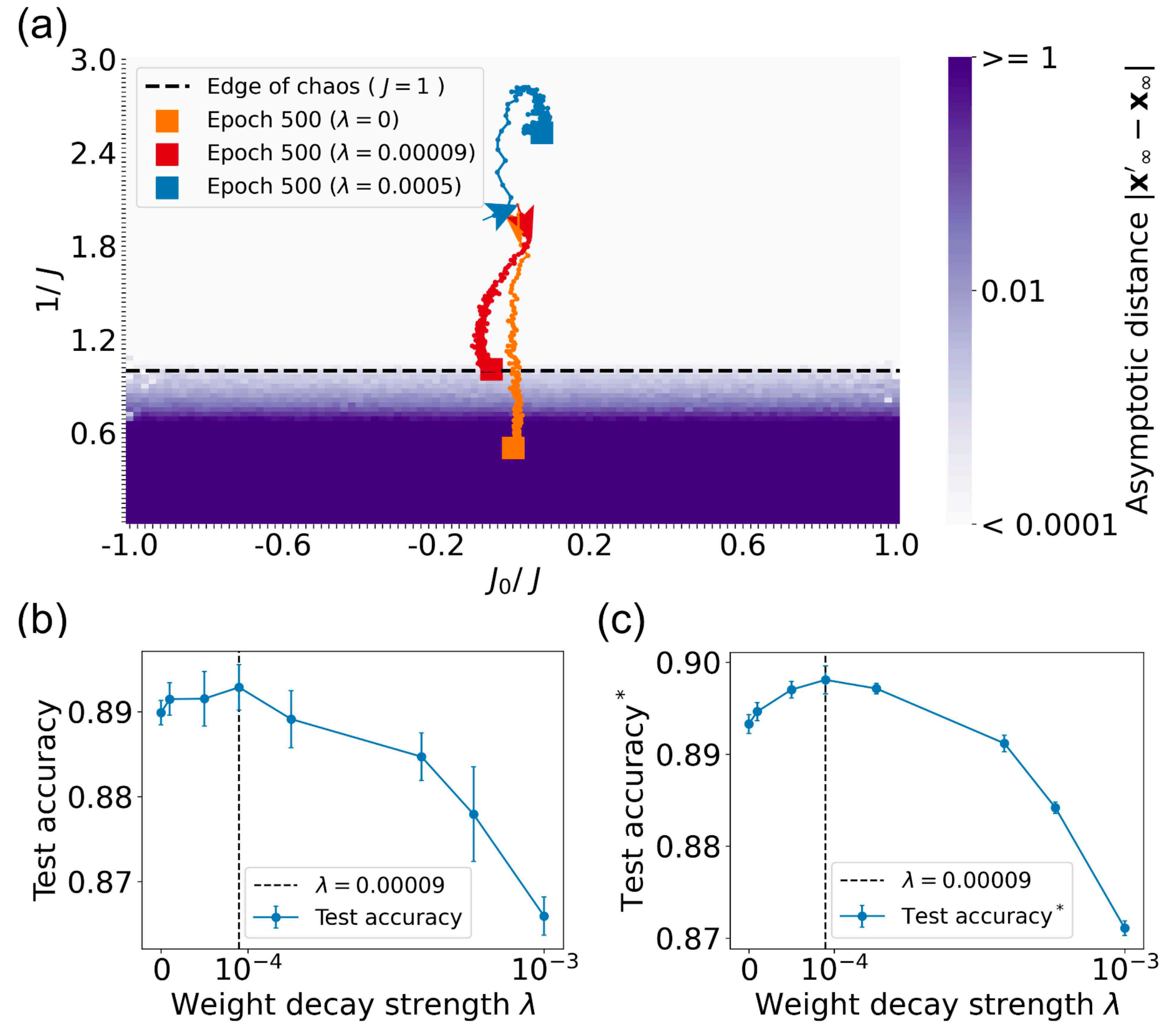

Fig 3: Model evolution path in the order-chaos phase diagram and related learning curves. Source: [2]

Conclusion

This work presents some theoretical guidance for regularization in Neural Networks. The training process with SGD is found to move the model from order to chaos, and the model has the highest generalization performance at the transition between these two phases. Regularizing the model to the edge of chaos lets the model achieve a balance between overfitting and underfitting [2].

References

[1] Bak, P. (1996). How Nature Works: The Science of Self-Organized Criticality. Copernicus.
[2] Zhang, L., Feng, L., Chen, K., & Lai, C. H. (2021). Edge of chaos as a guiding principle for modern neural network training. arXiv preprint arXiv:2107.09437.
[3] Yang, Y., Hodgkinson, L., Theisen, R., Zou, J., Gonzalez, J. E., Ramchandran, K., & Mahoney, M. W. (2021). *Taxonomizing local versus global structure in neural network loss landscapes.* arXiv preprint arXiv:2107.11228.
[4] Shwartz-Ziv, R., & Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. arXiv preprint arXiv:1703.00810.
[5] Baity-Jesi, M., Sagun, L., Geiger, M., Spigler, S., Ben Arous, G., Cammarota, C., ... Biroli, G. (2019). *Comparing dynamics: deep neural networks versus glassy systems. Journal of Statistical Mechanics: Theory and Experiment, 2019*(12), 124013. doi:10.1088/1742-5468/ab3281